REF1 MODIFIED PLANTS AND PLANT SEEDS

The present application was funded in part with government support under grant number de-fg02-94er20138 from the Department of Energy. The government may have certain rights in this invention.

The present Application claims priority to U.S. Provisional Application Serial Number 60/468,608, filed May 7, 2003.

FIELD OF THE INVENTION

The present invention relates to plant cell wall synthesis, and to plant secondary metabolite synthesis. In particular, the present invention provides plant aldehyde dehydrogenases with a novel capacity of oxidizing hydroxycinnamaldehydes to hydroxycinnamic acids, and genes encoding the enzyme, as well as variants of the gene, and methods of using the genes and variants of the gene to modify plant chemistry.

BACKGROUND OF THE INVENTION

One major feature which distinguishes plant cells from animal cells is the presence of a cell wall in plant cells. The cell wall is a non-living, semi-rigid or rigid wall containing cellulose. Not only does the presence of a cell wall form the basis of many of the characteristics of plants as organisms, but it also affects many commercial and industrial uses of plants.

Cell walls typically contain two layers: a primary wall, and an intercellular substance or middle lamella. In addition, many cells deposit another wall layer, the secondary wall. Secondary walls are particularly important in specialized cells that have functions such as conduction and strengthening. The most obvious distinguishing feature of secondary walls in most plant cells is the incorporation of lignins, which are complex networks of aromatic compounds called phenylpropanoids. The phenylpropanoids, hydroxycinnamoyl alcohols, and "monolignols" (such as *p*-coumaryl, coniferyl, and sinapyl alcohols) account for most of the lignin networks. The monolignols are linked by way of ester, ether, or carbon-carbon bonds.

The characteristics of different plant cell walls greatly affect the quality of plant materials, and thus their commercial and industrial utility. Plant cell wall materials are used

in human and animal food, textiles, wood, paper, and medicines. For example, the cell walls of fruits and vegetables are now recognized as important dietary components, and may protect against cancer of the colon, coronary heart disease, diabetes, and other ailments. Fibers obtained from cotton, flax, ramie and sisal plants are used to produce textiles, while cellulosic cell wall components are used to form paper and cardboard.

Modification of various cell wall constituents is a goal in the food processing, agriculture, and biotechnology industries. Successful achievement of this goal depends on understanding the molecular basis for the mechanical, textural and chemical properties of plant-derived materials. Thus, investigations of enzymes involved in cell wall metabolism are steps towards the goals of producing crops with desired characteristics by enhancing commercially valuable traits or abolishing undesirable ones. For example, the pulp and paper industry, which processes trees into cellulose, and the livestock industry, which depends on the transformation of plant materials, including cell walls, into muscle tissue, are striving to reduce the lignin content in their respective sources of fiber and fodder. In the paper industry, reducing lignin content or altering lignin quality would reduce organochlorine wastes and cut costs tremendously, as the paper industry currently uses extractions to purify cellulose from wood. Cell wall bound ferulic acid cross-links the hemicelluose fraction, making cellulose extraction difficult. Moreover, if total lignin content could be reduced by diverting metabolic carbon flow away from lignin biosynthesis, considerable savings could be achieved. In the livestock industry, lignin-carbohydrate interactions exert a great influence on digestibility of forage crops by animals, and the kind of lignin present, rather than the total amount present, is often a crucial factor. For example, ferulic is a major cell-wall bound lignin monomer, and its presence reduces digestibility of forage crops. Thus, decreasing the concentration of ferulic acid would increase digestibility of forage crops.

Thus, there is a need to manipulate the lignin quantity and/or quality of plant cell walls. For example, plants with modified lignin could result in new forage crops that exhibit greater digestibility without sacrificing the strengthening function of lignin to the water-conducting cells of the plant.

SUMMARY OF THE INVENTION

Thus, it is an object of the present invention to provide methods of decreasing ferulic acid content of cell walls of forage plants, thus increasing their digestibility. It is also an object of the present invention to provide methods to modify the quantity and/or quality of

lignin in plant cell walls. It is a further objective of the present invention to provide methods to modify the quantity and/or quality of hydroxycinnamic acid content of plant cells. It is yet a further object of the present invention to provide compositions comprising isolated coding sequences for and polypeptides of a plant protein which is an aldehyde dehydrogenase with a novel capacity of oxidizing hydroxycinnamaldehydes to hydroxycinnamic acids, for example, with the ability to oxidize coniferaldehyde and sinapaldehyde to ferulic acid and sinapic acid, respectively. These and other objectives are met by the present invention.

In some embodiments, the present invention provides a method to reduce sinapine content in a plant seed, comprising decreasing expression of an *REF1* gene in a plant seed during development of the plant seed, thereby decreasing sinapine content of the plant seed. In other embodiments, the present invention provides a method to reduce sinapine content in a plant seed, comprising growing a plant comprising a heterologous nucleic acid sequence, wherein expression of the heterologous nucleic acid sequence is able to decrease activity of REF1 in a seed of the plant, under conditions effective to decrease sinapine content of the plant seed. In further embodiments, the heterologous nucleic acid sequence encodes a nucleic acid product which interferes with expression of an *REF1* gene in the plant seed, wherein the interference is based upon a coding sequence of the *REF1* gene, under conditions effective to decrease sinapine content in the plant seed. In any of these embodiments, the plant seed is a Brassica or a *Sinapis alba* plant seed; in further embodiments, the plant is a *Brassica napus* plant seed; and in yet further embodiments, the plant is a canola plant seed.

In other embodiments, the present invention provides a method to decrease hydroxycinnamic acid content in a plant cell, comprising decreasing expression of an *REF1* gene in a plant cell, thereby decreasing sinapine content of the plant cell. In some embodiments, the present invention provides a method to decrease hydroxycinnamic acid content in a plant cell, comprising growing a plant cell comprising a heterologous nucleic acid sequence, wherein expression of the heterologous nucleic acid sequence is able to decrease activity of REF1 in the plant cell, under conditions effective to decrease hydroxycinnamic acid content of the plant cell. In further embodiments, the heterologous nucleic acid sequence encodes a nucleic acid product which interferes with expression of an *REF1* gene, wherein the interference is based upon a coding sequence of the *REF1* gene, under conditions effective to decrease hydroxycinnamic acid content in the plant cell.

In other embodiments, the present invention provides a method to decrease hydroxycinnamic acid content in a plant cell, comprising treating a plant cell such that the treatment decreases expression of an *REF1* gene in the plant cell, and growing the treated plant cell under conditions effective to decrease hydroxycinnamic acid content of the cell. In further embodiments, the treatment comprises transfecting the plant cell with a heterologous nucleic acid, wherein the heterologous nucleic acid is able to decrease expression of the *REF1* gene in the plant cell.

In yet other embodiments, the present invention provides a method to increase digestibility of plant forage or animal feed, comprising growing a plant treated to decrease expression of an *REF1* gene under conditions effective to decrease hydroxycinnamic acid content of cells in the plant, wherein the plant is a forage plant or is used as animal feed. In some further embodiments, the treatment comprises transfecting the plant with a heterologous nucleic acid sequence, wherein expression of the heterologous nucleic acid sequence is able to decrease activity of REF1 in cells of the plant. In yet further embodiments, the heterologous nucleic acid sequence encodes a nucleic acid product which interferes with expression of an *REF1* gene, wherein the interference is based upon a coding sequence of the *REF1* gene. In other embodiments, the hydroxycinnamic acid is ferulic acid. In yet other embodiments, the plant is a grass or legume. In other embodiments, present invention provides a method to increase digestibility of plant forage or animal feed, comprising transfecting a plant cell with a heterologous nucleic acid sequence, wherein expression of the heterologous nucleic acid sequence is able to decrease activity of REF1 in the cell, and growing a plant from the plant cell under conditions effective to decrease hydroxycinnamic acid content of cells in the plant, wherein the plant is a forage plant or is used as animal feed. In yet further embodiments, the heterologous nucleic acid sequence encodes a nucleic acid product which interferes with expression of an *REF1* gene, wherein the interference is based upon a coding sequence of the *REF1* gene. In other embodiments, the hydroxycinnamic acid is ferulic acid. In yet other embodiments, the plant is a grass or legume.

In other embodiments, the present invention provides a method to modify a lignin property in a plant cell, a plant tissue, a plant organ, a plant seed, or a plant, comprising growing a plant cell, a plant tissue, a plant organ, a plant seed, or a plant comprising a heterologous nucleic acid sequence, wherein the heterologous nucleic acid sequence is able to modulate activity of REF1 in the plant cell, the plant tissue, the plant organ, the plant

seed, or the plant, under conditions effective to modify a lignin property in the plant cell, the plant tissue, the plant organ, the plant seed, or the plant.

In some further embodiments, the heterologous nucleic acid sequence is able to increase expression of an *REF1* gene in the plant cell, the plant tissue, the plant organ, the plant seed, or the plant. In any of these embodiments, the heterologous nucleic acid sequence encodes an Arabidopsis REF1; in further embodiments, the heterologous nucleic acid sequence encodes SEQ ID NO:3. In other of these embodiments, the heterologous nucleic acid sequence encodes a homolog of SEQ ID NO:3; in yet other of these embodiments, the heterologous nucleic aid sequence encodes an REF1 homolog sequence shown in Table 1. In yet other further embodiments, lignin levels are decreased in the plant cell, the plant tissue, the plant organ, the plant seed, or the plant.
In yet other further embodiments, the heterologous nucleic acid sequence is able to decrease expression of the *REF1* gene in the plant cell, the plant tissue, the plant organ, the plant seed, or the plant.

In certain embodiments, the present invention provides transgenic plants or seeds that comprise a heterologous gene encoding a mutant *REF1* gene with decreased REF1 activity. In particular embodiments, the present invention provides transgenic plants or seeds that comprise a gene encoding a mutant *REF1* gene with decreased REF1 activity, wherein the mutant *REF1* gene is linked to a heterologous promoter (e.g., a promoter not naturally found in the plant that causes high levels of expression of the mutant *REF1* gene). In some embodiments, the present invention provides transgenic plants or seeds that express a variant/mutant form of the *REF1* gene with decreased REF1 activity. In additional embodiments, the transgenic plants or seeds do not express wild-type REF1, or express wild-type REF1 at reduced levels (e.g. reduced levels compared to wild-type expression levels).

In some embodiments, the present invention provides plants or plant seeds comprising a heterologous nucleic acid sequence, wherein the heterologous nucleic acid sequence, or the expression of the nucleic acid sequence, decreases activity of REF1 in the plant or the plant seed. In other embodiments, the heterologous nucleic acid sequence, or the expression of the heterologous nucleic acid sequence, results in reduced sinapine content in the plant or the plant seed. In particular embodiments, the heterologous nucleic acid sequence, or the expression of the heterologous nucleic acid sequence, results in reduced hydroxycinnamic acid content in the plant or the plant seed.

In other embodiments, the present invention provides methods comprising: a) treating a plant or plant seed such that treatment decreases expression of an *REF1* gene in the plant or plant seed; and b) growing the treated plant or plant seed. In particular embodiments, the growing results in a plant with reduced sinapine content. In further embodiments, the growing results in a plant with reduced hydroxycinnamic acid content.

DESCRIPTION OF THE FIGURES

Figure 1 shows the revised phenylpropanoid pathway in Arabidopsis that leads to the biosynthesis of lignin precursors from phenylalanine.

Figure 2 shows positional cloning of the REF1 gene. The REF1 gene was mapped to a 154 kb region between markers CER 435669 and g4711 using a mapping population of 550 F2 plants. This region is covered by three P1 clones containing 39 annotated genes. No recombinants were found with marker CER438851 located on the P1 clone MOB24 suggesting a very tight linkage of the REF1 gene to this marker. A putative aldehyde dehydrogenase gene located less than 8 kb from the CER43851 marker was identified as the REF1 gene by mutant complementation and sequence analysis.

Figure 3 shows a schematic representation of the intron-exon organization of the Arabidopsis REF1 gene.

Figure 4 shows an *REF1* EST sequence (SEQ ID NO:1, panel A), an REF1 open reading frame sequence (SEQ ID NO:2, panel B), and an REF1 protein sequence (SEQ ID NO:3, panel C).

Figure 5 shows exemplary sequences of sequences highly homologous to REF1 from other plants.

Figure 6 shows the results of complementation analysis of *ref1* mutants.

A schematic representation of the genomic DNA sequence used for the complementation of *ref1* plants is shown in panel A. A 7414 bp *Kpn* I - *Xho* I DNA fragment of the *REF1* gene comprising of 3297 bp of the promoter region, 3298 bp of the coding region and 819 bp of the 3'-UTR was ligated to *Kpn* I and *Xho* I digested pBluescript to generate pCC619. A Kpn I - Pst I fragment from pCC619 containing the 7414 bp *REF1* DNA was subcloned into a *Kpn* I and *Pst* I digested pCAMBIA2300 vector to generate plasmid pCC621. Plasmid pCC623 was constructed by subcloning a *Bam* HI fragment from pCC619 containing 1390 bp of the *REF1* promoter region, 3298 bp of the coding region and 819 bp of the 3'-UTR into a Bam HI digested pCAMBIA2300 vector.

The sinapoylmalate content in the leaves of the wild-type columbia and in the ref1-1, ref1-2 and ref1-4 plants transformed with pCAMBIA2300, pCC621, pCC623 vectors is shown in panel B. Error bar represents the standard deviation of sinapoylmalate content determined from three independent transgenic or wild-type plants.

Figure 7 shows aldehyde dehydrogenase activity in WT and *ref1* mutants. Sinapaldehyde dehydrogenase activity (panel A) and coniferaldehyde dehydrogenase activity (panel B) was determined using desalted protein extracts that were isolated from 21 day old WT and *ref1* seedlings. Error bars represents standard deviation determined from three enzyme assays.

DEFINITIONS

To facilitate an understanding of the present invention, a number of terms and phrases as used herein are defined below:

The term "plant" is used in it broadest sense. It includes, but is not limited to, any species of woody, ornamental or decorative, crop or cereal, fruit or vegetable plant, and photosynthetic green algae (*for example, Chlamydomonas reinhardtii*). It also refers to a plurality of plant cells which are largely differentiated into a structure that is present at any stage of a plant's development. Such structures include, but are not limited to, a fruit, shoot, stem, leaf, flower petal, etc. The term "plant tissue" includes differentiated and undifferentiated tissues of plants including those present in roots, shoots, leaves, pollen, seeds and tumors, as well as cells in culture (*for example*, single cells, protoplasts, embryos, callus, etc.). Plant tissue may be in planta, in organ culture, tissue culture, or cell culture. The term "plant part" as used herein refers to a plant structure or a plant tissue.

The term "crop" or "crop plant" is used in its broadest sense. The term includes, but is not limited to, any species of plant or algae edible by humans or used as a feed for animals or used, or consumed by humans, or any plant or algae used in industry or commerce.

The term plant cell "compartments or organelles" is used in its broadest sense. The term includes but is not limited to, the endoplasmic reticulum, Golgi apparatus, trans Golgi network, plastids, sarcoplasmic reticulum, glyoxysomes, mitochondrial, chloroplast, and nuclear membranes, and the like.

The term “host cell” refers to any cell capable of replicating and/or transcribing and/or translating a heterologous gene.

The term "aldehyde dehydrogenase (ALDH)" refers to an NAD(P)+ enzyme that oxidizes a substrate selected from the group including a wide spectrum of endogenous and exogenous aliphatic and aromatic aldehydes. Based upon the standardized ALDH nomenclature system proposed recently (Vasiliou et al. (1999) Pharmacogenetics 9(4): 421-434), REF1 is a member of the ALDH2 family; in particular, REF1 is an ALDH2C2.

Recommended nomenclature of eukaryotic ALDH genes has been proposed based on divergent evolution and chromosomal mapping (Vasiliou, V et al. (1999) Pharmacogenetics 9(4): 421-434). In this proposal, known ALDHs are divided into a total of 18 families (comprising 37 subfamilies). An ALDH protein from one gene family is defined as having approximately greater than or equal to 40% amino acid identity to that from another family. Two members of the same subfamily exhibit approximately greater than or equal to 60% amino acid identity, and are expected to be located at the same subchromosomal site. For naming each gene, it is proposed that the root symbol 'ALDH' denoting 'aldehyde dehydrogenase' be followed by an Arabic number representing the family and, when needed, a letter designating the subfamily and an Arabic number denoting the individual gene within the subfamily.

The term "REF1" refers to an aldehyde dehydrogenase with activity toward hydroxycinnamaldehydes, or a "hydroxycinnamaldehyde dehydrogenase." REF1 oxidizes both sinapaldehyde and coniferaldehyde to their respective acids, sinapic acid and ferulic acid. The oxidation of sinapaldehyde and coniferaldehyde to their respective acids is referred to as sinapaldehyde dehydrogenase (SALDH) and coniferaldehyde dehydrogenase (CALDH) activity, respectively. Thus, an REF1 has SALDH and/or CALDH activity; the ratio of SALDH to CALDH activity may vary among different REF1 enzymes.

An REF1 has been identified from Arabidopsis (see Fig. 4). The amino acid sequence of the Arabidopsis REF1 (AtREF1) is used as a reference sequence by which to identify other REF1 sequences, or REF1 homologs, both in Arabidopsis and in other plants. Exemplary homologs are shown in Fig. 5, and described in Table 1. An REF1 has at least about 45% amino acid identity with AtREF1, and SALDH and/or CALDH activity; typically, an REF1 has at least about 50%, or about 55%, or about 60% amino acid identity with AtREF1.

The term "REF1 activity" refers to enzymatic activity of an REF1 polypeptide. REF1 activity may be increased or decreased or otherwise modified. For example, REF1 activity may be decreased by decreasing the amount of REF1 polypeptide in a plant cell; such a decrease may occur as a result of decreased expression of an *REF1* gene in a plant

cell. Expression of an *REF1* plant gene may be decreased by any method which results the presence of less REF1 polypeptide in a cell, as for example by the presence of anti-sense coding sequences. REF1 activity may also be decreased by modifying the polypeptide in some fashion, such that, for example, the amount of product produced in a unit of time decreases, or the stability of the polypeptide decreases.

The term "ferulate 5-hydroxylase" or "F5H" refers to a hydroxylase with activity toward coniferaldehyde and coniferyl alcohol as well as toward ferulic acid, where the affinity of F5H for coniferaldehyde and coniferyl alcohol is much greater than its affinity for ferulic acid. By much greater it is meant that the affinity is at least about 10 fold, preferably at least about 50 fold, more preferably at least about 100 fold, and more preferably from about 100 to 1000 fold and higher.

The term "caffeic acid/5-hydroxyferulic acid O-methyltransferase" or "COMT" refers to a methyl transferase which is capable of converting 5-hydroxyconiferaldehyde and 5-hydroxyconiferyl alcohol to sinapaldehyde and sinapyl alcohol, respectively.

The term "lignin property" refers to a characteristic of lignin that can be modified by a change of expression of an *REF1* gene in a plant cell. Lignins are the most obvious distinguishing feature of secondary walls in most plant cells, and comprise complex networks of aromatic compounds called phenylpropanoids. The phenylpropanoids, hydroxycinnamoyl alcohols, and "monolignols" (such as *p*-coumaryl, coniferyl, and sinapyl alcohols) account for most of the lignin networks. The monolignols are linked by way of ester, ether, or carbon-carbon bonds. Lignin properties that can be modified include but are not limited to the amount of total lignin, or the amount of a specific lignin class, or the composition of the lignin.

The term "competes for binding" is used in reference to a first polypeptide with enzymatic activity which binds to the same substrate as does a second polypeptide with enzymatic activity, where the second polypeptide is variant of the first polypeptide or a related or dissimilar polypeptide. The efficiency (*for example*, kinetics or thermodynamics) of binding by the first polypeptide may be the same as or greater than or less than the efficiency substrate binding by the second polypeptide. For example, the equilibrium binding constant (K_D) for binding to the substrate may be different for the two polypeptides.

The terms "protein" and "polypeptide" refer to compounds comprising amino acids joined via peptide bonds and are used interchangeably.

As used herein, where "amino acid sequence" is recited herein to refer to an amino acid sequence of a protein molecule, "amino acid sequence" and like terms, such as

"polypeptide" or "protein" are not meant to limit the amino acid sequence to the complete, native amino acid sequence associated with the recited protein molecule; furthermore, an "amino acid sequence" can be deduced from the nucleic acid sequence encoding the protein.

The term "portion" when used in reference to a protein (as in "a portion of a given protein") refers to fragments of that protein. The fragments may range in size from four amino acid residues to the entire amino sequence minus one amino acid.

The term "chimera" when used in reference to a polypeptide refers to the expression product of two or more coding sequences obtained from different genes, that have been cloned together and that, after translation, act as a single polypeptide sequence. Chimeric polypeptides are also referred to as "hybrid" polypeptides. The coding sequences includes those obtained from the same or from different species of organisms.

The term "fusion" when used in reference to a polypeptide refers to a chimeric protein containing a protein of interest joined to an exogenous protein fragment (the fusion partner). The fusion partner may serve various functions, including enhancement of solubility of the polypeptide of interest, as well as providing an "affinity tag" to allow purification of the recombinant fusion polypeptide from a host cell or from a supernatant or from both. If desired, the fusion partner may be removed from the protein of interest after or during purification.

The term "homolog" or "homologous" when used in reference to a polypeptide refers to a high degree of sequence identity between two polypeptides, or to a high degree of similarity between the three-dimensional structure or to a high degree of similarity between the active site and the mechanism of action.

The term "homology" when used in relation to amino acid sequences refers to a degree of similarity or identity. There may be partial homology or complete homology (*in other words*, identity). "Sequence identity" refers to a measure of relatedness between two or more amino acid sequences, and is given as a percentage with reference to the total comparison length. The identity calculation takes into account those amino acid residues that are identical and in the same relative positions in their respective larger sequences. Calculations of identity may be performed by algorithms contained within computer programs. In a preferred embodiment, a homolog has a greater than 60% sequence identity, and more preferable greater than 75% sequence identity, and still more preferably greater than 90% sequence identity, with a reference sequence.

The terms "variant" and "mutant" when used in reference to a polypeptide refer to an amino acid sequence that differs by one or more amino acids from another, usually related

polypeptide. The variant may have "conservative" changes, wherein a substituted amino acid has similar structural or chemical properties (*for example*, replacement of leucine with isoleucine). More rarely, a variant may have "non-conservative" changes (*for example*, replacement of a glycine with a tryptophan). Similar minor variations may also include amino acid deletions or insertions (*in other words* additions), or both. Guidance in determining which and how many amino acid residues may be substituted, inserted or deleted without abolishing biological activity may be found using computer programs well known in the art, for example, DNAStar software. Variants can be tested in functional assays. Preferred variants have less than 10%, and preferably less than 5%, and still more preferably less than 2% changes (whether substitutions, deletions, and so on).

The term "gene" refers to a nucleic acid (*for example*, DNA or RNA) sequence that comprises coding sequences necessary for the production of an RNA, or a polypeptide or its precursor (*for example*, proinsulin). A functional polypeptide can be encoded by a full length coding sequence or by any portion of the coding sequence as long as the desired activity or functional properties (*for example*, enzymatic activity, ligand binding, signal transduction, etc.) of the polypeptide are retained. The term "portion" when used in reference to a gene refers to fragments of that gene. The fragments may range in size from a few nucleotides to the entire gene sequence minus one nucleotide. Thus, "a nucleotide comprising at least a portion of a gene" may comprise fragments of the gene or the entire gene.

The term "gene" also encompasses the coding regions of a structural gene and includes sequences located adjacent to the coding region on both the 5' and 3' ends for a distance of about 1 kb on either end such that the gene corresponds to the length of the full-length mRNA. The sequences which are located 5' of the coding region and which are present on the mRNA are referred to as 5' non-translated sequences. The sequences which are located 3' or downstream of the coding region and which are present on the mRNA are referred to as 3' non-translated sequences. The term "gene" encompasses both cDNA and genomic forms of a gene. A genomic form or clone of a gene contains the coding region interrupted with non-coding sequences termed "introns" or "intervening regions" or "intervening sequences." Introns are segments of a gene which are transcribed into nuclear RNA (hnRNA); introns may contain regulatory elements such as enhancers. Introns are removed or "spliced out" from the nuclear or primary transcript; introns therefore are absent in the messenger RNA (mRNA) transcript. The mRNA functions during translation to specify the sequence or order of amino acids in a nascent polypeptide.

In addition to containing introns, genomic forms of a gene may also include sequences located on both the 5' and 3' end of the sequences which are present on the RNA transcript. These sequences are referred to as "flanking" sequences or regions (these flanking sequences are located 5' or 3' to the non-translated sequences present on the mRNA transcript). The 5' flanking region may contain regulatory sequences such as promoters and enhancers which control or influence the transcription of the gene. The 3' flanking region may contain sequences which direct the termination of transcription, posttranscriptional cleavage and polyadenylation.

The term "heterologous gene" refers to a gene encoding a factor that is not in its natural environment (*in other words*, has been altered by the hand of man). For example, a heterologous gene includes a gene from one species introduced into another species. A heterologous gene also includes a gene native to an organism that has been altered in some way (*for example*, mutated, added in multiple copies, linked to a non-native promoter or enhancer sequence, etc.). Heterologous genes may comprise plant gene sequences that comprise cDNA forms of a plant gene; the cDNA sequences may be expressed in either a sense (to produce mRNA) or anti-sense orientation (to produce an anti-sense RNA transcript that is complementary to the mRNA transcript). Heterologous genes are distinguished from endogenous plant genes in that the heterologous gene sequences are typically joined to nucleotide sequences comprising regulatory elements such as promoters that are not found naturally associated with the gene for the protein encoded by the heterologous gene or with plant gene sequences in the chromosome, or are associated with portions of the chromosome not found in nature (*for example*, genes expressed in loci where the gene is not normally expressed).

The term "oligonucleotide" refers to a molecule comprised of two or more deoxyribonucleotides or ribonucleotides, preferably more than three, and usually more than ten. The exact size will depend on many factors, which in turn depends on the ultimate function or use of the oligonucleotide. The oligonucleotide may be generated in any manner, including chemical synthesis, DNA replication, reverse transcription, or a combination thereof.

The term "an oligonucleotide having a nucleotide sequence encoding a gene" or "a nucleic acid sequence encoding" a specified polypeptide refers to a nucleic acid sequence comprising the coding region of a gene or in other words the nucleic acid sequence which encodes a gene product. The coding region may be present in either a cDNA, genomic DNA or RNA form. When present in a DNA form, the oligonucleotide may be single-

stranded (*in other words*, the sense strand) or double-stranded. Suitable control elements such as enhancers/promoters, splice junctions, polyadenylation signals, *etc.* may be placed in close proximity to the coding region of the gene if needed to permit proper initiation of transcription and/or correct processing of the primary RNA transcript. Alternatively, the coding region utilized in the expression vectors of the present invention may contain endogenous enhancers/promoters, splice junctions, intervening sequences, polyadenylation signals, *etc.* or a combination of both endogenous and exogenous control elements.

The terms "complementary" and "complementarity" refer to polynucleotides (*in other words,* a sequence of nucleotides) related by the base-pairing rules. For example, for the sequence "A-G-T," is complementary to the sequence "T-C-A." Complementarity may be "partial," in which only some of the nucleic acids' bases are matched according to the base pairing rules. Or, there may be "complete" or "total" complementarity between the nucleic acids. The degree of complementarity between nucleic acid strands has significant effects on the efficiency and strength of hybridization between nucleic acid strands. This is of particular importance in amplification reactions, as well as detection methods which depend upon binding between nucleic acids.

The term "homology" when used in relation to nucleic acids refers to a degree of complementarity. There may be partial homology or complete homology (*in other words*, identity). "Sequence identity" refers to a measure of relatedness between two or more nucleic acids, and is given as a percentage with reference to the total comparison length. The identity calculation takes into account those nucleotide residues that are identical and in the same relative positions in their respective larger sequences. Calculations of identity may be performed by algorithms contained within computer programs such as "GAP" (Genetics Computer Group, Madison, Wis.) and "ALIGN" (DNAStar, Madison, Wis.). A partially complementary sequence is one that at least partially inhibits (or competes with) a completely complementary sequence from hybridizing to a target nucleic acid is referred to using the functional term "substantially homologous." The inhibition of hybridization of the completely complementary sequence to the target sequence may be examined using a hybridization assay (Southern or Northern blot, solution hybridization and the like) under conditions of low stringency. A substantially homologous sequence or probe will compete for and inhibit the binding (*in other words,* the hybridization) of a sequence which is completely homologous to a target under conditions of low stringency. This is not to say that conditions of low stringency are such that non-specific binding is permitted; low stringency conditions require that the binding of two sequences to one another be a specific

(*in other words*, selective) interaction. The absence of non-specific binding may be tested by the use of a second target which lacks even a partial degree of complementarity (*for example*, less than about 30% identity); in the absence of non-specific binding the probe will not hybridize to the second non-complementary target.

When used in reference to a double-stranded nucleic acid sequence such as a cDNA or genomic clone, the term "substantially homologous" refers to any probe which can hybridize to either or both strands of the double-stranded nucleic acid sequence under conditions of low stringency as described *infra*.

Low stringency conditions when used in reference to nucleic acid hybridization comprise conditions equivalent to binding or hybridization at 42EC in a solution consisting of 5X SSPE (43.8 g/l NaCl, 6.9 g/l $NaH_2PO_4XH_2O$ and 1.85 g/l EDTA, pH adjusted to 7.4 with NaOH), 0.1% SDS, 5X Denhardt's reagent [50X Denhardt's contains per 500 ml: 5 g Ficoll (Type 400, Pharmacia), 5 g BSA (Fraction V; Sigma)] and 100 :g/ml denatured salmon sperm DNA followed by washing in a solution comprising 5X SSPE, 0.1% SDS at 42EC when a probe of about 500 nucleotides in length is employed.

High stringency conditions when used in reference to nucleic acid hybridization comprise conditions equivalent to binding or hybridization at 42EC in a solution consisting of 5X SSPE (43.8 g/l NaCl, 6.9 g/l $NaH_2PO_4XH_2O$ and 1.85 g/l EDTA, pH adjusted to 7.4 with NaOH), 0.5% SDS, 5X Denhardt's reagent and 100 :g/ml denatured salmon sperm DNA followed by washing in a solution comprising 0.1X SSPE, 1.0% SDS at 42EC when a probe of about 500 nucleotides in length is employed.

It is well known that numerous equivalent conditions may be employed to comprise low stringency conditions; factors such as the length and nature (DNA, RNA, base composition) of the probe and nature of the target (DNA, RNA, base composition, present in solution or immobilized, etc.) and the concentration of the salts and other components (*for example*, the presence or absence of formamide, dextran sulfate, polyethylene glycol) are considered and the hybridization solution may be varied to generate conditions of low stringency hybridization different from, but equivalent to, the above listed conditions. In addition, the art knows conditions that promote hybridization under conditions of high stringency (*for example*, increasing the temperature of the hybridization and/or wash steps, the use of formamide in the hybridization solution, etc.).

When used in reference to a double-stranded nucleic acid sequence such as a cDNA or genomic clone, the term "substantially homologous" refers to any probe that can

hybridize to either or both strands of the double-stranded nucleic acid sequence under conditions of low to high stringency as described above.

When used in reference to a single-stranded nucleic acid sequence, the term "substantially homologous" refers to any probe that can hybridize (*in other words*, it is the complement of) the single-stranded nucleic acid sequence under conditions of low to high stringency as described above.

The term "hybridization" refers to the pairing of complementary nucleic acids. Hybridization and the strength of hybridization (*in other words*, the strength of the association between the nucleic acids) is impacted by such factors as the degree of complementary between the nucleic acids, stringency of the conditions involved, the T_m of the formed hybrid, and the G:C ratio within the nucleic acids. A single molecule that contains pairing of complementary nucleic acids within its structure is said to be "self-hybridized."

The term "T_m" refers to the "melting temperature" of a nucleic acid. The melting temperature is the temperature at which a population of double-stranded nucleic acid molecules becomes half dissociated into single strands. The equation for calculating the T_m of nucleic acids is well known in the art. As indicated by standard references, a simple estimate of the T_m value may be calculated by the equation: $T_m = 81.5 + 0.41(\% G + C)$, when a nucleic acid is in aqueous solution at 1 M NaCl (*See for example*, Anderson and Young, Quantitative Filter Hybridization (1985) *in Nucleic Acid Hybridization*). Other references include more sophisticated computations that take structural as well as sequence characteristics into account for the calculation of T_m.

As used herein the term "stringency" refers to the conditions of temperature, ionic strength, and the presence of other compounds such as organic solvents, under which nucleic acid hybridizations are conducted. With "high stringency" conditions, nucleic acid base pairing will occur only between nucleic acid fragments that have a high frequency of complementary base sequences. Thus, conditions of "low" stringency are often required with nucleic acids that are derived from organisms that are genetically diverse, as the frequency of complementary sequences is usually less.

"Amplification" is a special case of nucleic acid replication involving template specificity. It is to be contrasted with non-specific template replication (*in other words*, replication that is template-dependent but not dependent on a specific template). Template specificity is here distinguished from fidelity of replication (*in other words*, synthesis of the proper polynucleotide sequence) and nucleotide (ribo- or deoxyribo-) specificity. Template

specificity is frequently described in terms of "target" specificity. Target sequences are "targets" in the sense that they are sought to be sorted out from other nucleic acid. Amplification techniques have been designed primarily for this sorting out.

Template specificity is achieved in most amplification techniques by the choice of enzyme. Amplification enzymes are enzymes that, under conditions they are used, will process only specific sequences of nucleic acid in a heterogeneous mixture of nucleic acid. For example, in the case of Q ∃replicase, MDV-1 RNA is the specific template for the replicase (Kacian *et al.* (1972) Proc. Natl. Acad. Sci. USA, 69:3038). Other nucleic acid will not be replicated by this amplification enzyme. Similarly, in the case of T7 RNA polymerase, this amplification enzyme has a stringent specificity for its own promoters (Chamberlin *et al.* (1970) Nature, 228:227). In the case of T4 DNA ligase, the enzyme will not ligate the two oligonucleotides or polynucleotides, where there is a mismatch between the oligonucleotide or polynucleotide substrate and the template at the ligation junction (Wu and Wallace (1989) Genomics, 4:560). Finally, *Taq* and *Pfu* polymerases, by virtue of their ability to function at high temperature, are found to display high specificity for the sequences bounded and thus defined by the primers; the high temperature results in thermodynamic conditions that favor primer hybridization with the target sequences and not hybridization with non-target sequences (H.A. Erlich (ed.) (1989) *PCR Technology*, Stockton Press).

The term "amplifiable nucleic acid" refers to nucleic acids that may be amplified by any amplification method. It is contemplated that "amplifiable nucleic acid" will usually comprise "sample template."

The term "sample template" refers to nucleic acid originating from a sample that is analyzed for the presence of "target" (defined below). In contrast, "background template" is used in reference to nucleic acid other than sample template that may or may not be present in a sample. Background template is most often inadvertent. It may be the result of carryover, or it may be due to the presence of nucleic acid contaminants sought to be purified away from the sample. For example, nucleic acids from organisms other than those to be detected may be present as background in a test sample.

The term "primer" refers to an oligonucleotide, whether occurring naturally as in a purified restriction digest or produced synthetically, which is capable of acting as a point of initiation of synthesis when placed under conditions in which synthesis of a primer extension product which is complementary to a nucleic acid strand is induced, (*in other words*, in the presence of nucleotides and an inducing agent such as DNA polymerase and at

a suitable temperature and pH). The primer is preferably single stranded for maximum efficiency in amplification, but may alternatively be double stranded. If double stranded, the primer is first treated to separate its strands before being used to prepare extension products. Preferably, the primer is an oligodeoxyribonucleotide. The primer must be sufficiently long to prime the synthesis of extension products in the presence of the inducing agent. The exact lengths of the primers will depend on many factors, including temperature, source of primer and the use of the method.

The term "polymerase chain reaction" ("PCR") refers to the method of K.B. Mullis U.S. Patent Nos. 4,683,195, 4,683,202, and 4,965,188, that describe a method for increasing the concentration of a segment of a target sequence in a mixture of genomic DNA without cloning or purification. This process for amplifying the target sequence consists of introducing a large excess of two oligonucleotide primers to the DNA mixture containing the desired target sequence, followed by a precise sequence of thermal cycling in the presence of a DNA polymerase. The two primers are complementary to their respective strands of the double stranded target sequence. To effect amplification, the mixture is denatured and the primers then annealed to their complementary sequences within the target molecule. Following annealing, the primers are extended with a polymerase so as to form a new pair of complementary strands. The steps of denaturation, primer annealing, and polymerase extension can be repeated many times (*in other words*, denaturation, annealing and extension constitute one "cycle"; there can be numerous "cycles") to obtain a high concentration of an amplified segment of the desired target sequence. The length of the amplified segment of the desired target sequence is determined by the relative positions of the primers with respect to each other, and therefore, this length is a controllable parameter. By virtue of the repeating aspect of the process, the method is referred to as the "polymerase chain reaction" (hereinafter "PCR"). Because the desired amplified segments of the target sequence become the predominant sequences (in terms of concentration) in the mixture, they are said to be "PCR amplified."

With PCR, it is possible to amplify a single copy of a specific target sequence in genomic DNA to a level detectable by several different methodologies (*for example*, hybridization with a labeled probe; incorporation of biotinylated primers followed by avidin-enzyme conjugate detection; incorporation of ^{32}P-labeled deoxynucleotide triphosphates, such as dCTP or dATP, into the amplified segment). In addition to genomic DNA, any oligonucleotide or polynucleotide sequence can be amplified with the appropriate

set of primer molecules. In particular, the amplified segments created by the PCR process itself are, themselves, efficient templates for subsequent PCR amplifications.

The terms "PCR product," "PCR fragment," and "amplification product" refer to the resultant mixture of compounds after two or more cycles of the PCR steps of denaturation, annealing and extension are complete. These terms encompass the case where there has been amplification of one or more segments of one or more target sequences.

The term "amplification reagents" refers to those reagents (deoxyribonucleotide triphosphates, buffer, etc.), needed for amplification except for primers, nucleic acid template, and the amplification enzyme. Typically, amplification reagents along with other reaction components are placed and contained in a reaction vessel (test tube, microwell, etc.).

The term "reverse-transcriptase" or "RT-PCR" refers to a type of PCR where the starting material is mRNA. The starting mRNA is enzymatically converted to complementary DNA or "cDNA" using a reverse transcriptase enzyme. The cDNA is then used as a "template" for a "PCR" reaction.

The terms "gene expression" or "expression of a gene" or the like refer to the process of converting genetic information encoded in a gene into RNA (*for example*, mRNA, rRNA, tRNA, or snRNA) through "transcription" of the gene (*in other words*, via the enzymatic action of an RNA polymerase), and/or into protein, through "translation" of mRNA. Gene expression can be regulated, manipulated, or modified, at many stages in the process. For example, gene expression products can be increased or decreased when compared to the presence or amount of the gene expression products in the absence of regulation, manipulation, or modification. There are many means of regulating, manipulating, or modifying gene expression. These include but are not limited to "up-regulation" or "activation," and "down-regulation" or "repression." "Up-regulation" or "activation" refers to regulation that increases the production of gene expression products (*for example*, RNA or protein), while "down-regulation" or "repression" refers to regulation that decrease production. Molecules (*for example*, transcription factors) that are involved in up-regulation or down-regulation are often called "activators" and "repressors," respectively.

The terms "in operable combination", "in operable order" and "operably linked" refer to the linkage of nucleic acid sequences in such a manner that a nucleic acid molecule capable of directing the transcription of a given gene and/or the synthesis of a desired

protein molecule is produced. The term also refers to the linkage of amino acid sequences in such a manner so that a functional protein is produced.

The term "regulatory element" refers to a genetic element which controls some aspect of the expression of nucleic acid sequences. For example, a promoter is a regulatory element which facilitates the initiation of transcription of an operably linked coding region. Other regulatory elements are splicing signals, polyadenylation signals, termination signals, *etc.*

Transcriptional control signals in eukaryotes comprise "promoter" and "enhancer" elements. Promoters and enhancers consist of short arrays of DNA sequences that interact specifically with cellular proteins involved in transcription (Maniatis, *et al.*, Science 236:1237, 1987). Promoter and enhancer elements have been isolated from a variety of eukaryotic sources including genes in yeast, insect, mammalian and plant cells. Promoter and enhancer elements have also been isolated from viruses and analogous control elements, such as promoters, are also found in prokaryotes. The selection of a particular promoter and enhancer depends on the cell type used to express the protein of interest. Some eukaryotic promoters and enhancers have a broad host range while others are functional in a limited subset of cell types (*for review, see* Voss, *et al.*, Trends Biochem. Sci., 11:287, 1986; and Maniatis, *et al., supra* 1987).

The terms "promoter element," "promoter," or "promoter sequence" as used herein, refer to a DNA sequence that is located at the 5' end (*in other words* precedes) the protein coding region of a DNA polymer. The location of most promoters known in nature precedes the transcribed region. The promoter functions as a switch, activating the expression of a gene. If the gene is activated, it is said to be transcribed, or participating in transcription. Transcription involves the synthesis of mRNA from the gene. The promoter, therefore, serves as a transcriptional regulatory element and also provides a site for initiation of transcription of the gene into mRNA.

Promoters may be tissue specific or cell specific. The term "tissue specific" as it applies to a promoter refers to a promoter that is capable of directing selective expression of a nucleotide sequence of interest to a specific type of tissue (*for example*, seeds) in the relative absence of expression of the same nucleotide sequence of interest in a different type of tissue (*for example*, leaves). Tissue specificity of a promoter may be evaluated by, for example, operably linking a reporter gene to the promoter sequence to generate a reporter construct, introducing the reporter construct into the genome of a plant such that the reporter construct is integrated into every tissue of the resulting transgenic plant, and detecting the

expression of the reporter gene (*for example*, detecting mRNA, protein, or the activity of a protein encoded by the reporter gene) in different tissues of the transgenic plant. The detection of a greater level of expression of the reporter gene in one or more tissues relative to the level of expression of the reporter gene in other tissues shows that the promoter is specific for the tissues in which greater levels of expression are detected. The term "cell type specific" as applied to a promoter refers to a promoter which is capable of directing selective expression of a nucleotide sequence of interest in a specific type of cell in the relative absence of expression of the same nucleotide sequence of interest in a different type of cell within the same tissue. The term "cell type specific" when applied to a promoter also means a promoter capable of promoting selective expression of a nucleotide sequence of interest in a region within a single tissue. Cell type specificity of a promoter may be assessed using methods well known in the art, *for example*, immunohistochemical staining. Briefly, tissue sections are embedded in paraffin, and paraffin sections are reacted with a primary antibody which is specific for the polypeptide product encoded by the nucleotide sequence of interest whose expression is controlled by the promoter. A labeled (*for example*, peroxidase conjugated) secondary antibody which is specific for the primary antibody is allowed to bind to the sectioned tissue and specific binding detected (*for example*, with avidin/biotin) by microscopy.

Promoters may be constitutive or regulatable. The term "constitutive" when made in reference to a promoter means that the promoter is capable of directing transcription of an operably linked nucleic acid sequence in the absence of a stimulus (*for example*, heat shock, chemicals, light, *etc.*). Typically, constitutive promoters are capable of directing expression of a transgene in substantially any cell and any tissue. Exemplary constitutive plant promoters include, but are not limited to SD Cauliflower Mosaic Virus (CaMV SD; *see for example*, U.S. Pat. No. 5,352,605, incorporated herein by reference), mannopine synthase, octopine synthase (ocs), superpromoter (*see for example*, WO 95/14098), and *ubi3* (*see for example*, Garbarino and Belknap (1994) Plant Mol. Biol. 24:119-127) promoters. Such promoters have been used successfully to direct the expression of heterologous nucleic acid sequences in transformed plant tissue.

In contrast, a "regulatable" promoter is one which is capable of directing a level of transcription of an operably linked nuclei acid sequence in the presence of a stimulus (*for example*, heat shock, chemicals, light, *etc.*) which is different from the level of transcription of the operably linked nucleic acid sequence in the absence of the stimulus.

The enhancer and/or promoter may be "endogenous" or "exogenous" or "heterologous." An "endogenous" enhancer or promoter is one that is naturally linked with a given gene in the genome. An "exogenous" or "heterologous" enhancer or promoter is one that is placed in juxtaposition to a gene by means of genetic manipulation (*in other words*, molecular biological techniques) such that transcription of the gene is directed by the linked enhancer or promoter. For example, an endogenous promoter in operable combination with a first gene can be isolated, removed, and placed in operable combination with a second gene, thereby making it a "heterologous promoter" in operable combination with the second gene. A variety of such combinations are contemplated (*for example*, the first and second genes can be from the same species, or from different species.

The presence of "splicing signals" on an expression vector often results in higher levels of expression of the recombinant transcript in eukaryotic host cells. Splicing signals mediate the removal of introns from the primary RNA transcript and consist of a splice donor and acceptor site (Sambrook, *et al.* (1989) Molecular Cloning: A Laboratory Manual, 2nd ed., Cold Spring Harbor Laboratory Press, New York, pp. 16.7-16.8). A commonly used splice donor and acceptor site is the splice junction from the 16S RNA of SV40.

Efficient expression of recombinant DNA sequences in eukaryotic cells requires expression of signals directing the efficient termination and polyadenylation of the resulting transcript. Transcription termination signals are generally found downstream of the polyadenylation signal and are a few hundred nucleotides in length. The term "poly(A) site" or "poly(A) sequence" as used herein denotes a DNA sequence which directs both the termination and polyadenylation of the nascent RNA transcript. Efficient polyadenylation of the recombinant transcript is desirable, as transcripts lacking a poly(A) tail are unstable and are rapidly degraded. The poly(A) signal utilized in an expression vector may be "heterologous" or "endogenous." An endogenous poly(A) signal is one that is found naturally at the 3' end of the coding region of a given gene in the genome. A heterologous poly(A) signal is one which has been isolated from one gene and positioned 3' to another gene. A commonly used heterologous poly(A) signal is the SV40 poly(A) signal. The SV40 poly(A) signal is contained on a 237 bp *Bam*HI/*Bcl*I restriction fragment and directs both termination and polyadenylation (Sambrook, *supra*, at 16.6-16.7).

The term "selectable marker" refers to a gene which encodes an enzyme having an activity that confers resistance to an antibiotic or drug upon the cell in which the selectable marker is expressed, or which confers expression of a trait which can be detected (*for example.*, luminescence or fluorescence). Selectable markers may be "positive" or

"negative." Examples of positive selectable markers include the neomycin phosphotransferase (NPTII) gene which confers resistance to G418 and to kanamycin, and the bacterial hygromycin phosphotransferase gene (*hyg*), which confers resistance to the antibiotic hygromycin. Negative selectable markers encode an enzymatic activity whose expression is cytotoxic to the cell when grown in an appropriate selective medium. For example, the HSV-*tk* gene is commonly used as a negative selectable marker. Expression of the HSV-*tk* gene in cells grown in the presence of gancyclovir or acyclovir is cytotoxic; thus, growth of cells in selective medium containing gancyclovir or acyclovir selects against cells capable of expressing a functional HSV TK enzyme.

The term "vector refers to nucleic acid molecules that transfer DNA segment(s) from one cell to another. The term "vehicle" is sometimes used interchangeably with "vector."

The terms "expression vector" or "expression cassette" refer to a recombinant DNA molecule containing a desired coding sequence and appropriate nucleic acid sequences necessary for the expression of the operably linked coding sequence in a particular host organism. Nucleic acid sequences necessary for expression in prokaryotes usually include a promoter, an operator (optional), and a ribosome binding site, often along with other sequences. Eukaryotic cells are known to utilize promoters, enhancers, and termination and polyadenylation signals.

The term "transfection" refers to the introduction of foreign DNA into cells. Transfection may be accomplished by a variety of means known to the art including calcium phosphate-DNA co-precipitation, DEAE-dextran-mediated transfection, polybrene-mediated transfection, glass beads, electroporation, microinjection, liposome fusion, lipofection, protoplast fusion, viral infection, biolistics (*for example*, particle bombardment) and the like.

The terms "infecting" and "infection" when used with a bacterium refer to co-incubation of a target biological sample, (*for example*, cell, tissue, *etc.*) with the bacterium under conditions such that nucleic acid sequences contained within the bacterium are introduced into one or more cells of the target biological sample.

The term "*Agrobacterium*" refers to a soil-borne, Gram-negative, rod-shaped phytopathogenic bacterium which causes crown gall. The term "*Agrobacterium*" includes, but is not limited to, the strains *Agrobacterium tumefaciens*, (which typically causes crown gall in infected plants), and *Agrobacterium rhizogens* (which causes hairy root disease in infected host plants). Infection of a plant cell with *Agrobacterium* generally results in the production of opines (*for example*, nopaline, agropine, octopine etc.) by the infected cell.

Thus, *Agrobacterium* strains which cause production of nopaline (*for example*, strain LBA4301, C58, A208, GV3101) are referred to as "nopaline-type" *Agrobacteria*; *Agrobacterium* strains which cause production of octopine (*for example*, strain LBA4404, Ach5, B6) are referred to as "octopine-type" *Agrobacteria*; and *Agrobacterium* strains which cause production of agropine (*for example*, strain EHA105, EHA101, A281) are referred to as "agropine-type" *Agrobacteria*.

The terms "bombarding, "bombardment," and "biolistic bombardment" refer to the process of accelerating particles towards a target biological sample (*for example*, cell, tissue, *etc.*) to effect wounding of the cell membrane of a cell in the target biological sample and/or entry of the particles into the target biological sample. Methods for biolistic bombardment are known in the art (*for example*, U.S. Patent No. 5,584,807, the contents of which are incorporated herein by reference), and are commercially available (*for example*, the helium gas-driven microprojectile accelerator (PDS-1000/He, BioRad).

The term "microwounding" when made in reference to plant tissue refers to the introduction of microscopic wounds in that tissue. Microwounding may be achieved by, for example, particle bombardment as described herein.

The term "transgenic" when used in reference to a plant or fruit or seed (*in other words*, a "transgenic plant" or "transgenic fruit" or a "transgenic seed") refers to a plant or fruit or seed that contains at least one heterologous gene in one or more of its cells. The term "transgenic plant material" refers broadly to a plant, a plant structure, a plant tissue, a plant seed or a plant cell that contains at least one heterologous gene in one or more of its cells.

The terms "transformants" or "transformed cells" include the primary transformed cell and cultures derived from that cell without regard to the number of transfers. All progeny may not be precisely identical in DNA content, due to deliberate or inadvertent mutations. Mutant progeny that have the same functionality as screened for in the originally transformed cell are included in the definition of transformants.

The term "wild-type" when made in reference to a gene refers to a gene which has the characteristics of a gene isolated from a naturally occurring source. The term "wild-type" when made in reference to a gene product refers to a gene product which has the characteristics of a gene product isolated from a naturally occurring source. A wild-type gene is that which is most frequently observed in a population and is thus arbitrarily designated the "normal" or "wild-type" form of the gene. In contrast, the term "modified" or "mutant" when made in reference to a gene or to a gene product refers, respectively, to a

gene or to a gene product which displays modifications in sequence and/or functional properties (*in other words*, altered characteristics) when compared to the wild-type gene or gene product. It is noted that naturally-occurring mutants can be isolated; these are identified by the fact that they have altered characteristics when compared to the wild-type gene or gene product.

The term "antisense" refers to a deoxyribonucleotide sequence whose sequence of deoxyribonucleotide residues is in reverse 5' to 3' orientation in relation to the sequence of deoxyribonucleotide residues in a sense strand of a DNA duplex. A "sense strand" of a DNA duplex refers to a strand in a DNA duplex which is transcribed by a cell in its natural state into a "sense mRNA." Thus an "antisense" sequence is a sequence having the same sequence as the non-coding strand in a DNA duplex. The term "antisense RNA" refers to a RNA transcript that is complementary to all or part of a target primary transcript or mRNA and that blocks the expression of a target gene by interfering with the processing, transport and/or translation of its primary transcript or mRNA. The complementarity of an antisense RNA may be with any part of the specific gene transcript, *in other words*, at the 5' non-coding sequence, 3' non-coding sequence, introns, or the coding sequence. In addition, as used herein, antisense RNA may contain regions of ribozyme sequences that increase the efficacy of antisense RNA to block gene expression. "Ribozyme" refers to a catalytic RNA and includes sequence-specific endoribonucleases. "Antisense inhibition" refers to the production of antisense RNA transcripts capable of preventing the expression of the target protein.

The term "siRNAs" refers to short interfering RNAs. In some embodiments, siRNAs comprise a duplex, or double-stranded region, of about 18-25 nucleotides long; often siRNAs contain from about two to four unpaired nucleotides at the 3' end of each strand. At least one strand of the duplex or double-stranded region of a siRNA is substantially homologous to or substantially complementary to a target RNA molecule. The strand complementary to a target RNA molecule is the "antisense strand;" the strand homologous to the target RNA molecule is the "sense strand," and is also complementary to the siRNA antisense strand. siRNAs may also contain additional sequences; non-limiting examples of such sequences include linking sequences, or loops, which link the two strands of the double strand, as well as stem and other folded structures, which may be present within the linking sequence. siRNAs appear to function as key intermediaries in triggering RNA interference in invertebrates and in vertebrates, and in triggering sequence-specific RNA degradation during posttranscriptional gene silencing in plants.

The term "target RNA molecule" refers to an RNA molecule to which at least one strand of the short double-stranded region of an siRNA is homologous or complementary. Typically, when such homology or complementary is about 100%, the siRNA is able to silence or inhibit expression of the target RNA molecule. Although it is believed that processed mRNA is a target of siRNA, the present invention is not limited to any particular hypothesis, and such hypotheses are not necessary to practice the present invention. Thus, it is contemplated that other RNA molecules may also be targets of siRNA. Such targets include unprocessed mRNA, ribosomal RNA, and viral RNA genomes.

The term "RNA interference" or "RNAi" refers to the silencing or decreasing of gene expression by siRNAs. It is the process of sequence-specific, post-transcriptional gene silencing in animals and plants, initiated by siRNA that is homologous in its duplex region to the sequence of the silenced gene. The gene may be endogenous or exogenous to the organism, present integrated into a chromosome or present in a transfection vector which is not integrated into the genome. The expression of the gene is either completely or partially inhibited. RNAi may also be considered to inhibit the function of a target RNA; the function of the target RNA may be complete or partial.

The term "interferes with expression of a gene" or the like refers to generally to silencing or decreasing of the gene expression. Both antisense RNA and interfering RNA are examples of molecules which interfere with expression of a gene.

The term "posttranscriptional gene silencing" or "PTGS" refers to silencing of gene expression in plants after transcription, and appears to involve the specific degradation of mRNAs synthesized from gene repeats.

The term "overexpression" refers to the production of a gene product in transgenic organisms that exceeds levels of production in normal or non-transformed organisms. The term "cosuppression" refers to the expression of a foreign gene which has substantial homology to an endogenous gene resulting in the suppression of expression of both the foreign and the endogenous gene. The term "altered levels" refers to the production of gene product(s) in transgenic organisms in amounts or proportions that differ from that of normal or non-transformed organisms.

The term "recombinant" when made in reference to a nucleic acid molecule refers to a nucleic acid molecule which is comprised of segments of nucleic acid joined together by means of molecular biological techniques. The term "recombinant" when made in reference to a protein or a polypeptide refers to a protein molecule which is expressed using a recombinant nucleic acid molecule.

The terms "Southern blot analysis" and "Southern blot" and "Southern" refer to the analysis of DNA on agarose or acrylamide gels in which DNA is separated or fragmented according to size followed by transfer of the DNA from the gel to a solid support, such as nitrocellulose or a nylon membrane. The immobilized DNA is then exposed to a labeled probe to detect DNA species complementary to the probe used. The DNA may be cleaved with restriction enzymes prior to electrophoresis. Following electrophoresis, the DNA may be partially depurinated and denatured prior to or during transfer to the solid support. Southern blots are a standard tool of molecular biologists (J. Sambrook *et al.* (1989) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Press, NY, pp 9.31-9.58).

The term "Northern blot analysis" and "Northern blot" and "Northern" as used herein refer to the analysis of RNA by electrophoresis of RNA on agarose gels to fractionate the RNA according to size followed by transfer of the RNA from the gel to a solid support, such as nitrocellulose or a nylon membrane. The immobilized RNA is then probed with a labeled probe to detect RNA species complementary to the probe used. Northern blots are a standard tool of molecular biologists (J. Sambrook, *et al.* (1989) *supra*, pp 7.39-7.52).

The terms "Western blot analysis" and "Western blot" and "Western" refers to the analysis of protein(s) (or polypeptides) immobilized onto a support such as nitrocellulose or a membrane. A mixture comprising at least one protein is first separated on an acrylamide gel, and the separated proteins are then transferred from the gel to a solid support, such as nitrocellulose or a nylon membrane. The immobilized proteins are exposed to at least one antibody with reactivity against at least one antigen of interest. The bound antibodies may be detected by various methods, including the use of radiolabeled antibodies.

The term "isolated" when used in relation to a nucleic acid, as in "an isolated oligonucleotide" refers to a nucleic acid sequence that is identified and separated from at least one contaminant nucleic acid with which it is ordinarily associated in its natural source. Isolated nucleic acid is present in a form or setting that is different from that in which it is found in nature. In contrast, non-isolated nucleic acids, such as DNA and RNA, are found in the state they exist in nature. For example, a given DNA sequence (*for example*, a gene) is found on the host cell chromosome in proximity to neighboring genes; RNA sequences, such as a specific mRNA sequence encoding a specific protein, are found in the cell as a mixture with numerous other mRNA s which encode a multitude of proteins. However, isolated nucleic acid encoding a plant CPA-FAS includes, by way of example, such nucleic acid in cells ordinarily expressing a DES, where the nucleic acid is in a chromosomal location different from that of natural cells, or is otherwise flanked by a

different nucleic acid sequence than that found in nature. The isolated nucleic acid or oligonucleotide may be present in single-stranded or double-stranded form. When an isolated nucleic acid or oligonucleotide is to be utilized to express a protein, the oligonucleotide will contain at a minimum the sense or coding strand (*in other words*, the oligonucleotide may single-stranded), but may contain both the sense and anti-sense strands (*in other words*, the oligonucleotide may be double-stranded).

The term "purified" refers to molecules, either nucleic or amino acid sequences, that are removed from their natural environment, isolated or separated. An "isolated nucleic acid sequence" is therefore a purified nucleic acid sequence. "Substantially purified" molecules are at least 60% free, preferably at least 75% free, and more preferably at least 90% free from other components with which they are naturally associated. The term "purified" or "to purify" also refer to the removal of contaminants from a sample. The removal of contaminating proteins results in an increase in the percent of polypeptide of interest in the sample. In another example, recombinant polypeptides are expressed in plant, bacterial, yeast, or mammalian host cells and the polypeptides are purified by the removal of host cell proteins; the percent of recombinant polypeptides is thereby increased in the sample.

The term "sample" is used in its broadest sense. In one sense it can refer to a plant cell or tissue. In another sense, it is meant to include a specimen or culture obtained from any source, as well as biological and environmental samples. Biological samples may be obtained from plants or animals (including humans) and encompass fluids, solids, tissues, and gases. Environmental samples include environmental material such as surface matter, soil, water, and industrial samples. These examples are not to be construed as limiting the sample types applicable to the present invention.

DESCRIPTION OF THE INVENTION

The present invention relates to plant cell wall synthesis, and to the synthesis of soluble compounds in plants commonly known as secondary metabolites or natural products. In particular, the present invention provides plant aldehyde dehydrogenases with a novel capacity of oxidizing hydroxycinnamaldehydes to hydroxycinnamic acids, and genes encoding the enzymes, as well as variants of the genes, and methods of using the genes and variants of the gene to modify plant chemistry. The description below provides specific, but not limiting, illustrative examples of embodiments of the present invention.

The following description describes the development of the invention, including the

discovery of a plant aldehyde dehydrogenase with a novel capacity of oxidizing hydroxycinnamaldehydes to hydroxycinnamic acids, and genes encoding the enzyme from Arabidopsis, variants of the genes, and use of the genes to alter plant chemistry, including but not limited to altering cell wall constituents. The Arabidopsis gene is used to identify other plant sequences encoding enzymes with similar functions in other plants; these other enzymes are also used to alter plant chemistry.

I. Discovery of an Aldehyde Dehydrogenase that Converts Hydroxycinnamaldehydes into Hydroxycinnamic Acids

Plants produce a number of important secondary metabolites from phenylalanine via the phenylpropanoid pathway. These include lignin, flavonoids, isoflavonoids, lignins, tannins, salicylic acid and hydroxycinnamic acid esters. These metabolites serve a variety of functions in plants, which include but are not limited to mechanical strength, pollen viability, pest deterrence, disease resistance, UV protection, biotic and abiotic interactions (see, for example, Nicholson and Hammerschmidt, 1992; van der Meer et al., 1992; Chapple, 1994; Landry et al., 1995; Dixon et al., 1996; Dempsey et al., 1999).

Arabidopsis and other members of the Brassicaceae accumulate hydroxycinnamic acid esters, for example, sinapoylmalate in leaves, and sinapoylcholine in seeds (Chapple et al., 1992; Nair et al., 2000). Sinapoylmalate, a UV fluorescent compound, accumulates in the upper leaf epidermis and causes the wild-type Arabidopsis leaves to appear blue-green under UV light. In contrast, the sinapoylmalate deficient *fah1* mutant appears red under UV-light due to the fluorescence of chlorophyll (Chapple et al., 1992; Meyer et al., 1996). This fluorescent property of sinapoylmalate was used to identify a series of Arabidopsis mutants (*ref1*-8) that showed reduced epidermal fluorescence (*ref*); these mutants thus appeared to be perturbed in the phenylpropanoid pathway. These mutants have been utilized to better understand the phenylpropanoid pathway, resulting in revision of the conventional model of the pathway (Ruegger and Chapple, 2001; Franke *et al.*, 2002a; Franke *et al.*, 2002b; Hemm et al., 2003).

The conventional model of the phenylpropanoid pathway includes a series of hydroxylation and methylation reactions in which cinnamic acid derived from phenylalanine is converted into a variety of hydroxycinnamic acids that act as precursors for flavonoids, hydroxycinnamic acid conjugates, and lignin (Higuchi, 1981). Activation of these acids to their corresponding CoA thioesters, followed by two successive reductions, was then thought to provide the monolignol building blocks for the lignin polymer. This

simple model was challenged following the identification of caffeoyl CoA 3-O-methyltransferase (CCoAOMT) as part of an elicitor-induced plant defense response (Pakusch et al., 1989; Kühnl et al., 1989; Pakusch et al., 1991; Schmitt et al., 1991). Subsequent work suggested that this enzyme also functions in a route to lignin monomers in which *p*-coumaroyl CoA is converted to feruloyl CoA, thus providing an alternative to the conventional pathway involving free acids (Ye et al., 1994; Ye et al., 1995). The significance of this pathway was demonstrated by CCoAOMT down-regulation in transgenic tobacco and alfalfa. Reduced CCoAOMT activity was associated with reduced syringyl and guaiacyl monomer production and dramatically lowered lignin content (Zhong et al., 1998; Guo et al., 2001). These experiments provided the first indications that not all the hydroxycinnamic acids were obligatory players in lignin biosynthesis.

The precursors of guaiacyl and syringyl lignin in angiosperms are coniferyl alcohol and sinapyl alcohol, respectively. Ferulate 5-hydroxylase (F5H) is required for syringyl monomer biosynthesis, and for many years its role in lignin biosynthesis was thought to be the conversion of ferulic acid to 5-hydroxyferulic acid, the precursor to sinapic acid and syringyl lignin. More recently, two pieces of evidence suggested that F5H might function differently in syringyl lignin biosynthesis. First, when F5H-overexpressing Arabidopsis, tobacco, and poplar plants were analyzed, they were found to deposit lignin composed almost entirely of syringyl monomers (Meyer et al., 1998; Franke et al., 2000). Since a model of the lignin biosynthetic pathway that incorporates CCoAOMT provides an alternative route to guaiacyl monomers that would bypass F5H, it was difficult to rationalize how F5H overexpression could have such dramatic phenotypic effects if its *in vivo* substrate were actually ferulic acid. Second, labeled precursor feeding experiments in magnolia implicated coniferyl alcohol in sinapyl alcohol biosynthesis (Chen et al., 1999). These issues were resolved by heterologous expression and subsequent enzymatic analysis of F5H. Kinetic analysis of the Arabidopsis enzyme revealed that although F5H catalyzed the 5-hydroxylation of ferulic acid, it displayed a thousand-fold greater affinity for coniferaldehyde and coniferyl alcohol (Humphreys et al., 1999). Concurrent research demonstrated that recombinant sweetgum F5H exhibited a high affinity for coniferaldehyde, a substrate that also inhibited ferulate 5-hydroxylation (Osakabe et al., 1999). Taken together, this work suggests that in lignin biosynthesis, F5H likely acts on coniferaldehyde and coniferyl alcohol, downstream in the phenylpropanoid pathway from its previously described substrate, ferulic acid.

The identification of novel F5H activities immediately led to a search for an *O*-methyltransferase capable of converting the "new" F5H products into syringyl-substituted monomers. Although caffeic acid / 5-hydroxyferulic acid *O*-methyltransferase (COMT) was initially thought to be a bifunctional enzyme that used hydroxycinnamic acids as substrates (Davin and Lewis, 1992), recombinant COMT from Arabidopsis, sweetgum, poplar, and alfalfa have now all been demonstrated to have activity toward the novel F5H products (Humphreys et al., 1999; Osakabe et al., 1999; Li et al., 2000; Parvathi et al., 2001). Further, the K_m of recombinant poplar COMT for 5-hydroxyconiferaldehyde is five-fold lower than for 5-hydroxyferulate and almost 30-fold lower than for caffeic acid (Li et al., 2000). Similarly, a recent extensive study of recombinant alfalfa COMT found that, in comparison to seven other possible substrates in the phenylpropanoid pathway, COMT exhibited the highest $V_{max} K_m^{-1}$ values for 5-hydroxyconiferaldehyde and the lowest for caffeic acid (Parvathi et al., 2001). These *in vitro* data suggest that COMT acts after F5H in the conversion of coniferaldehyde and coniferyl alcohol to sinapaldehyde and sinapyl alcohol, respectively. This repositioning of COMT within the lignin biosynthetic pathway is also consistent with previous work with transgenic tobacco and poplar in which reduced COMT expression led to a marked reduction in syringyl lignin with little impact on total lignin (Atanassova et al., 1995; Van Doorsselaere et al., 1995).

In the new model of lignin biosynthesis (Humphreys and Chapple, 2002), COMT no longer has a function redundant with CCoAOMT, thus explaining the ability to generate lignin phenotypes by CCoAOMT down-regulation. If the optimal substrates for F5H and COMT are hydroxy-cinnamaldehyde and hydroxy-cinnamylalcohol derived substrates, then the question arises as to how ferulic and sinapic acid are synthesized in plants. It is possible that ferulic and sinapic acid are made by the traditional free acid pathway. However, the high K_m values exhibited by COMT, and particularly F5H for the necessary intermediates, do not make this an attractive theory. Several years ago it was postulated that sinapic acid may instead be made by the oxidation of sinapaldehyde (Humphreys et al., 1999). Previously it has been shown that *ref1* mutants (*ref1-1* to *ref1-7*) had reduced levels of sinapate esters in leaves and seeds but accumulated wild-type levels of lignin (Ruegger and Chapple, 2001). As described below, it was discovered that *ref1* mutants are defective in an aldehyde dehydrogenase gene. Moreover, it was surprisingly discovered that wild-type *REF1* gene encodes a functional aldehyde dehydrogenase that is capable of oxidizing both sinapaldehyde and coniferaldehyde to sinapic and ferulic acid, respectively.

Thus, although ferulic and sinapic were conventionally considered to be substrates for the formation of coniferaldehyde and sinapaldehyde, in fact, the pathway operates in the opposite direction. The revised phenylpropanoid pathway in Arabidopsis of the biosynthesis of lignin precursors from phenylalanine is shown in Fig. 1. This pathway is based upon current understanding of the pathway based upon the analysis the substrate specificity of F5H (Humphreys et al., 1999; Osakabe et al., 1999), the characterization of ref8 mutant (Franke et al., 2002a; Franke et al., 2002b), the work by Schoch et al. (2001) in which C3H was shown to hydroxylate p-coumaroyl shikimate and/or p-coumaroyl quinate, and the unexpected discovery of the function of REF1 as a hydroxycinnamaldehyde dehydrogenase.

Identification of REF1

Seven independent alleles of *ref1* were originally identified in a mutant screen (Ruegger and Chapple, 2001). All of the *ref1* mutants are morphologically similar to wild type, but exhibit varying levels of epidermal fluorescence when observed under UV light. HPLC analysis of the sinapate ester content in *ref1-1* and *ref1-2* mutants showed up to a 70% reduction in sinapoylmalate and sinapoylcholine content in leaves and seeds, respectively. In contrast, the lignin content and composition was comparable to that of wild-type plants. These observations indicate that the *ref1* mutation affects a portion of the phenylpropanoid pathway required for sinapate ester biosynthesis, but not for the deposition of lignin.

Linkage analysis using a mapping population of 550 F2 plants showed that the *REF1* locus was positioned between CAPS markers CER435669 and g4711 on chromosome 3. These two markers define a 154 kb region containing 39 annotated genes (Fig. 2). A CAPS marker between these two markers, CER438851, was in the zero recombination well, indicating that it is tightly linked to the *REF1* locus. The fully sequenced Arabidopsis genome was utilized to identify candidate genes within this mapping region. One of these genes was annotated as a putative aldehyde dehydrogenase (ALDH) and was located less than 8 kb from the CAPS marker CER438851 at which no recombinants had been identified. To test the hypothesis that defects in this putative ALDH gene were responsible for the *ref1* phenotype, the *ref1-3* allele, which had been generated by T-DNA mutagenesis, was used. Although previous attempts using plasmid rescue to isolate the *REF1* gene with this allele were unsuccessful, it was postulated that this allele might have been the result of a partial/abortive T-DNA insertion event that would still be

detectable by PCR. Indeed, PCR amplification using gene specific and T-DNA border primers showed that the putative ALDH gene was disrupted by a T-DNA insertion. Agarose gel electrophoresis of PCR products were obtained using primers as shown in Fig. 3. The REF1 gene is not amplified from the mutant DNA using flanking primers, presumably due to the size of the T-DNA insertion. Primers CC414 and CC415 were used to amplify a known PCR product as a positive control for the presence and quality of the DNA prepared from the mutant. The position of the T-DNA insertion is downstream of the sequences corresponding to primers CC587 and CC588. This finding has subsequently been verified by sequence analysis of a cloned border sequence. A schematic representation of the intron-exon organization of the Arabidopsis REF1 gene is shown in Fig. 3, along with the *ref1-3* T-DNA tagged mutant, and the primers used to amplify the gene.

To determine whether this gene was also affected in the other *ref1* alleles, RNA gel blot analysis was then performed. 5 μg of leaf total RNA from mutant and wild-type plants was hybridized with 32P-labelled REF1 DNA and cyclophilin (ROC3) DNA (as a loading control). The data were then plotted as a histogram representing the hybridization signal of REF1 mRNAs normalized against cyclophilin. The wild-type signal value was set as the reference (100%). In four of the six EMS-induced alleles, the level of gene expression was reduced to less than 60% of wild type. Given that transcript levels of alleles containing frameshift or nonsense mutations are frequently decreased (Hilleren and Parker, 1999), these data further suggested that the putative ALDH gene is a strong candidate for *REF1*.

To provide definitive evidence for this hypothesis, the putative ALDH genomic DNA corresponding to all of the EMS-induced alleles was sequenced. Sequence analysis showed that three of the *ref1* alleles had point mutations in the gene resulting in premature translational stop codons (*ref1-1* and *ref1-4*, Q344STOP; *ref1-2*, W452STOP). A mis-sense mutation is present in *ref1-6* (G416R) and in *ref1-7* (G152E) and there is a nucleotide change in the intron 6 / exon 7 splice site junction (agGT → aaGT) that may lead to improper splicing in the *ref1-5* allele.

An expressed sequence tag (EST) (gene bank # AA395226; clone ID 118C14XP) that was annotated to have the 5'-end of the *REF1* gene was obtained and sequenced completely; the sequence results are shown in Fig. 4. The 1625 bp cDNA sequence (SEQ ID NO:1, Fig. 4A) contained a 43 bp- 5' untranslated region (UTR), 1506 bp of the predicted open reading frame (ORF) (SEQ ID NO:2, Fig. 4B) and a 76 bp of the 3'-UTR. The deduced polypeptide consists of 501 amino acids (SEQ ID NO:3, Fig. 4C) with a predicted

molecular mass of 54.3 kD. Its sequence contains the conserved residues that have been identified in human liver ALDH as active site amino acids (Ni et al., 1997; Sheikh et al., 1997). The lack of N-or C-terminal extensions compared to the mammalian liver ALDH suggests that the REF1 is likely to be tetrameric. Based upon the standardized ALDH nomenclature system proposed recently (Vasiliou et al., 1999), REF1 is a member of the ALDH2 family of aldehyde dehydrogenases.

The AtREF1 amino acid sequence exhibits homology to other Arabidopsis aldehyde dehydrogenases, including betaine aldehyde dehydrogenase, methylmalonate semialdehyde dehydrogenase, and an ABA-inducible aldehyde dehydrogenase (Kirch et al., 2001). Interestingly, REF1 is more closely related (53% to 77% identity) to ALDH sequences annotated as of unknown function from *Medicago truncatula*, *Glycine max*, *Oryza sativa*, *Zea mays*, and *Nicotiana tabacum*; exemplary sequences are shown in Fig 5. The identification of more highly related sequences in other species indicates that *REF1* is not be unique to Arabidopsis; the elucidation of REF1 function will shed light on phenylpropanoid metabolism in a wide range of plant species, including those that are agronomically important, and allow manipulation of the resulting lignocelullosic products derived from them, as well as their profile of secondary metabolites.

To demonstrate that the *ref1* mutation is the result of a lesion in this aldehyde dehydrogenase gene, the *ref1* mutant was complemented with wild-type *REF1* gene. Two plasmids, pCC621 and pCC623, were constructed with the *REF1* promoter DNA sequences, the *REF1* coding sequence and 3'-untranslated region (Fig. 6A). Plasmids pCC621 and pCC623 contained 3.3 kb and 1.4 kb of the promoter region, respectively. These plasmids were transformed into *ref1* mutants using the floral dip method (Clough and Bent, 1998). When transformed into plants, these sequences were able to completely restore wild-type levels of sinapoylmalate accumulation in *ref1-1*, *ref1-2* and *ref1-4* plants (Fig. 6B). These results provide conclusive evidence that a mutation in the *REF1* gene is responsible for the *ref1* phenotype.

To test the hypothesis that the *REF1* gene encodes an aldehyde dehydrogenase with activity toward hydroxycinnamaldehydes, aldehyde dehydrogenase activity towards coniferaldehyde and sinapaldehyde was measured in crude desalted plant enzyme extracts of wild-type and *ref1* leaf tissue. *ref1* plants exhibited less than 20% of the aldehyde dehydrogenase activity found in wild-type extracts when both sinapaldehyde and coniferaldehyde were used as substrates. GC-MS analysis of the compounds present in the aldehyde dehydrogenase enzyme reaction confirmed production of both ferulic and sinapic

acid in the assays. These results strongly suggests that the Arabidopsis *REF1* gene encodes an aldehyde dehydrogenase that plays a major role in the synthesis of sinapic and ferulic acid.

As a final proof that the *REF1* gene encodes a hydroxycinnamaldehyde dehydrogenase, the *REF1* open reading frame was subcloned into pET30a protein expression vector and expressed in *E. coli*. Crude protein extracts from *E. coli* transformed with pET30a vector alone did not show any aldehyde dehydrogenase activity toward hydroxycinnamaldehyde substrates, whereas, *E. coli* transformed with pET30a-*REF1* were able to efficiently catalyze the oxidation of both sinapaldehyde and coniferaldehyde to sinapic acid and ferulic acid, respectively.

REF1 has unexpected activity

This finding that REF1 oxidizes both sinapaldehyde and coniferaldehyde was unexpected, since it is the opposite of the conventional model in which ferulic and sinapic acid are considered as precursors for the formation of coniferaldehyde and sinapaldehyde. These results also suggested that the *ref1* mutant may have as yet undetected phenotypes associated with ferulic acid biosynthesis. To evaluate whether the *ref1* mutation reduces the amount of cell wall bound ferulic acid, wild-type and *ref1* cell wall preparations were subjected to alkaline hydrolysis to release esterified phenolics. HPLC analysis of cell wall hydrolysates revealed that the *ref1* mutation resulted in reduced levels of ferulic acid in cell walls, to about half that of non-mutant plants. This suggests that the *REF1* gene plays a major role in the formation of cell wall linked ferulic acid, which is a major factor limiting the digestibility of maize and other crops used to feed livestock.

To determine whether the *ref1* mutation affects the accumulation of soluble ferulic acid derivatives, hydroxycinnamate ester accumulation in *fah1* plants was examined. Although the major phenolic ester that accumulates in the leaves of wild-type Arabidopsis is sinapoylmalate, it is absent in *fah1* mutants, and in its place, feruloylmalate accumulates in trace amounts. In a *ref1*-*fah1* double mutant, feruloylmalate content is reduced to less than 5 percent of that present in *fah1* plants. This suggests that REF1 activity is needed for the production of ferulic acid which is a precursor for feruloylmalate synthesis in the leaves of Arabidopsis.

CALDH and SALDH activity is Present in Different Plants

The results described above demonstrate that both coniferaldehyde dehydrogenase (CALDH) and sinapaldehyde dehydrogenase (SALDH) enzyme activities are present in Arabidopsis extracts. There are a wide range of plants that accumulate cell wall linked ferulate esters (Hatfield, 1993; Lam et al., 1996; Ralph et al., 1995). Thus, it was hypothesized that, based upon the results described above, the primary route to ferulic and sinapic acid synthesis in most plants is from coniferaldehyde or sinapaldehyde, respectively, by aldehyde dehydrogenase activity. This hypothesis was tested by analyzing enzyme extracts from different plants were analyzed for both CALDH and SALDH activity. The plants included representative samples of dicot (Arabidopsis, tobacco and radish), monocot (maize), gymnosperm (pine) and pteridophyte (fern). The results indicate that both CALDH and SALDH activity is present in all the plants analyzed (Example 2, Section G, Table 2). The specific activities of SALDH and CALDH varied among the different plants assayed. Wild type Arabidopsis enzyme extract had higher SALDH and CALDH specific activity than did the *ref1* plant extract. Conversely, the ratio of SALDH to CALDH activity was above 1 for wild-type plant extract and less than 1 for *ref1* plant extract, which imply that in *ref1* plants CALDH activity was predominant.

Radish, another member of the family Brassicaceae, has been reported to accumulate less sinapoylmalate and more feruloylmalate in leaves (Denault and Chapple, unpublished results). Interestingly, radish leaf extracts showed higher CALDH specific activity than SALDH, suggesting that the differences in amounts of sinapoylmalate and feruloylmalate in radish could reflect on the relative specificity of aldehyde dehydrogenases present in this plant. Tobacco leaf and fern frond extracts showed higher specific activity towards SALDH. Pine, which accumulates only coniferaldehyde derived phenylpropanoid products had very high specific activity towards coniferaldehyde. Pine xylem extracts had the lowest SALDH/CALDH ratio (0.2) as compared to other plant tissues analyzed. Maize leaf extracts also showed higher specific activity towards coniferaldehyde. These results indicate that both CALDH and SALDH activity are present in a wide range of plants.

II. Plant REF1 Polypeptides

The present invention provides compositions comprising purified REF1 polypeptides as well as compositions comprising variants of REF1, including homologs, mutants, fragments, and fusion proteins thereof (as described further below). These

compositions may be used as a dehydrogenase with for hydroxycinnamaldehyde substrates, by the methods of the present invention.

In some embodiments of the present invention, the polypeptide is a purified product, obtained from expression of a native gene in a cell, while in other embodiments it may be a product of chemical synthetic procedures, and in still other embodiments it may be produced by recombinant techniques using a prokaryotic or eukaryotic host (e.g., by bacterial, yeast, higher plant, insect and mammalian cells in culture). In some embodiments, depending upon the host employed in a recombinant production procedure, the polypeptide of the present invention may be glycosylated or may be non-glycosylated. In other embodiments, the polypeptides of the invention may also include an initial methionine amino acid residue.

A. Reaction Catalyzed

REF1 is an aldehyde dehydrogenase with activity toward hydroxycinnamaldehydes, or a hydroxycinnamaldehyde dehydrogenase. REF1 oxidizes both sinapaldehyde and coniferaldehyde to their respective acids, sinapic acid and ferulic acid. Thus, REF1 has coniferaldehyde dehydrogenase (CALDH) and sinapaldehyde dehydrogenase (SALDH) activity. The ratios of CALDH to SALDH may vary among REF1s from different plant sources; it is contemplated that the activity may range from about 100% CALDH activity to about 100% SALDH activity for any particular plant.

The activity of REF1 may be determined by either *in vivo* or *in vitro* assays. *In vivo* assays may be conducted by measuring endogenous ferulic acid and/or sinapic acid, or a product downstream from ferulic acid and/or sinapic acid, as for example by measuring sinapate esters, ferulate esters, sinapoylmalate, sinapylcholine, etc. as is described in Example 1, Section D. *In vitro* assays may be performed by the addition of exogenous substrates coniferaldehyde and/or sinapaldehyde, as for example is described in Example 1, Section E.

B. Arabidopsis REF1

In some embodiments, the polypeptide comprises a purified Arabidopsis REF1. In one embodiment, the polypeptide is encoded by the sequence shown in Fig. 4B (SEQ ID NO:2); in other embodiments, the polypeptide comprises the amino acid sequence shown in Fig. 4C (SEQ ID NO:3). The official Arabidopsis gene number or name for REF1 is At3g24503; the genomic sequence of REF1 identified by GENBANK® accession number

is AB020746 (co-ordinates 11108 to 14405), and the GENBANK® protein identification number is BAB01998.

Arabidopsis aldehyde dehydrogenases, denominated AtALDH1a and ALDH-2, have been reported (Skibbe, DS et al. (20020 Plant Molecular Biology 48: 751-764; US patent application Pub No.: US 2002/0162137 A1). AtALDH1a appears to be the same as REF1, as determined by a comparison of the deduced protein sequences of the genomic data from the Arabidopsis genome database and the BAC clone MIOB24 (gene bank accession # AB020746). However, the function of AtALDH1a in phenylpropanoid biosynthesis was not reported in either publication, nor was their potential utility in modifying phenylpropanoid metabolism identified. For example, Skibbe et al. reported that AtALDH1a was functionally characterized (as one of several cloned ALDH genes) and displayed ALDH activity, where acetaldehyde or glycloaldehyde were used as substrates. US patent application Pub No.: US 2002/0162137 A1 reported that ALDH-2 converts aldehyde to acetate. However, neither reference described the activity of the aldehyde dehydrogenases as involved in the phenylpropanoid pathway, much less as an aldehyde dehydrogenase with activity toward hydroxycinnamaldehydes (or a "hydroxycinnamaldehyde dehydrogenase"), as was discovered by the inventors. Moreover, neither reference described the ability of the aldehyde dehydrogenases as capable of oxidizing both sinapaldehyde and coniferaldehyde to their respective acids, sinapic acid and ferulic acid, as was discovered by the inventors.

C. REF1 from Other Plants

Moreover, the inventors have also identified REF1 sequences in other plants (see Table 1). These REF1 sequences have an amino acid sequence identity of at least 55% to that of REF1 from Arabidopsis (i.e., to SEQ ID NO:3) and are a "hydroxycinnamaldehyde dehydrogenase". For example, the closest heterologous genes for REF1 from dicots are about 75% identical, based upon amino acid similarity (see, for example, the sequence from Medicago, shown in Fig. 5), and those from monocots are about 65% identical (see, for example, the sequences from maize and rice, shown in Fig. 5).

Moreover, as noted above, the inventors have demonstrated that several other plants have both SALDH and CALDH activity. Some plants have been shown to have both REF1-like sequences (see Table 1) and SALDH and CALDH activity (see Table 2), for example, maize, and tobacco.

Thus, in other embodiments, the polypeptide comprises a purified REF1 obtained from another plant source. These plant sources include but are not limited to Arabidopsis, rice, maize, tobacco, barley, sorghum, soybean, wheat, alfalfa, and canola. In particular embodiments, these plant sources include *Medicago truncatula*, *Zea mays*, *Glycine max*, *Oryza sativa*, *Nicotiana tabacum*, *Hordeum vulgare*, *Triticum aestivum*, and Brassica species, and include the amino acid sequences listed in Table 1 and shown in Figure 5.

D. Variant REF1 Polypeptides

In other embodiments, the present invention provides isolated variants of the disclosed REF1 polypeptides; these variants include mutants, fragments, fusion proteins or functional equivalents of REF1. Exemplary variants are described further below.

E. Assay of REF1 Polypeptides

The activity of REF1 may be assayed in a number of ways. These include, but are not limited to, *in vivo* assays and *in vitro* assays, as described further below.

In some embodiments, enzyme activity is determined *in vivo* by complementing a mutant in which REF1 normally present is knocked out; thus, a nucleic acid sequence encoding an *REF1* gene is expressed in a transgenic organism, and the content and composition of an end product, such as sinapate esters, analyzed. Exemplary methods are described in the Examples.

In other embodiments, enzyme activity is determined *in vivo* by adding exogenous substrates to tissue samples obtained from an organism which may or may not be transgenic (transgenic organisms are described below). For example, in plants, tissue samples include but are not limited to leaf samples (such as discs), stem and root samples, and developing and mature seed embryonic or endosperm tissue. Typically, tissue samples are incubated with [^{14}C]coniferaldehyde substrate or [^{14}C]sinapaldehyde, which can be taken up and incorporated into ferulic acid and sinapic acid, respectively, and further downstream into soluble and cell wall-bound esters. Incubations generally proceed at room temperature in a buffered solution; the samples are then washed in buffer, and the labeled products extracted from tissue samples and separated and analyzed by HPLC, as for example, is described in the Examples.

In yet other embodiments, enzyme activity is determined *in vitro* in a cell-free homogenate or subcellular fraction obtained from an organism which may or may not be transgenic (transgenic organisms are described below), where the tissue is disrupted and

typically filtered or centrifuged to result in cell-free fractions. For example, in plants, subcellular fractions may be obtained from any of the types of tissues described above; the preparation of such fractions are well-known in the art. The subcellular fraction is then assayed for SALDH or CALDH activity, as for example is described in the Examples.

In yet other embodiments, enzyme activity is determined from an *in-vitro* nucleic acid expression system, to which a nucleic acid sequence having a coding sequence of the present invention (for example, encoding an REF1, as, for example, SEQ ID NO:3, or comprising an REF1 coding sequence, as, for example, SEQ ID NOs: 1or 2) is added and the encoded enzyme expressed, and the activity of the expressed enzyme determined. Such expression systems are well-known in the art, and include, for example reticulocyte lysate or wheat germ. The activity of newly-expressed enzyme is then analyzed as described above for cell-free homogenates or subcellular fractions.

F. Purification of REF1 Polypeptides

In some embodiments of the present invention, an REF1 polypeptide purified from organisms is provided; such organisms include transgenic organisms, comprising a heterologous *REF1* gene, as well as organisms in which REF1 occurs naturally. In other embodiments, an REF1 polypeptide is purified from an *in vitro* nucleic acid expression system, which comprises a nucleic acid sequence having a coding sequence of the present invention (for example, encoding an REF1, as, for example, SEQ ID NO:3, or comprising an REF1 coding sequence, as, for example, SEQ ID NOs: 1 or 3) and from which the expressed REF1 can be purified. The present invention provides a purified REF1 polypeptide as well as variants, including homologs, mutants, fragments, and fusion proteins thereof (as described further below).

The present invention also provides methods for recovering and purifying plant REF1 from an organism or from an *in vitro* nucleic acid expression system; exemplary organisms include single and multi-cellular organisms. When isolated from an organism, the cells are typically first disrupted and then fractionated before subsequent enzyme purification; disruption and fractionation methods are well-known.

Purification methods are also well-known, and include, but are not limited to, ammonium sulfate or ethanol precipitation, acid extraction, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, affinity chromatography, hydroxylapatite chromatography and lectin chromatography, and ioselectric focusing.

The present invention further provides nucleic acid sequences having a coding sequence of the present invention (*e.g.*, SEQ ID NOs: 1 or 2) fused in frame to a marker sequence that allows for expression alone or both expression and purification of the polypeptide of the present invention. A non-limiting example of a marker sequence is a hexahistidine tag that may be supplied by a vector, for example, a pQE-30 vector which adds a hexahistidine tag to the N terminal of an REF1 and which results in expression of the polypeptide in the case of a bacterial host, and in other embodiments by vector PT-23B, which adds a hexahistidine tag to the C terminal of an REF1 and which results in improved ease of purification of the polypeptide fused to the marker in the case of a bacterial host, or, for example, the marker sequence may be a hemagglutinin (HA) tag when a mammalian host is used. The HA tag corresponds to an epitope derived from the influenza hemagglutinin protein (Wilson *et al.* (1984) Cell 37:767).

G. Chemical Synthesis of REF1 Polypeptides

In some embodiments of the present invention, an REF1 protein is produced using chemical methods to synthesize either an entire REF1 amino acid sequence or a portion thereof. For example, peptides are synthesized by solid phase techniques, cleaved from the resin, and purified by preparative high performance liquid chromatography (See *e.g.*, Creighton (1983) Proteins Structures And Molecular Principles, W H Freeman and Co, New York N.Y.). In other embodiments of the present invention, the composition of the synthetic peptides is confirmed by amino acid analysis or sequencing (See *e.g.*, Creighton, supra).

Direct peptide synthesis can be performed using various solid-phase techniques (Roberge *et al.* (1995) Science, 269:202-204) and automated synthesis may be achieved, for example, using ABI 431A Peptide Synthesizer (Perkin Elmer) in accordance with the instructions provided by the manufacturer. Additionally, an amino acid sequence of an REF1, or any part thereof, may be altered during direct synthesis and/or combined using chemical methods with other sequences to produce a variant polypeptide.

H. Generation of REF1 Antibodies

In some embodiments of the present invention, antibodies are generated to allow for the detection and characterization of an REF1 protein. The antibodies may be prepared using various immunogens. In one embodiment, the immunogen is a REF1 peptide (*e.g.*, an amino acid sequence as depicted in SEQ ID NO:2, or fragments thereof) to generate

antibodies that recognize REF1. Such antibodies include, but are not limited to polyclonal, monoclonal, chimeric, single chain, Fab fragments, and Fab expression libraries.

Various procedures known in the art may be used for the production of polyclonal antibodies directed against an REF1. For the production of antibody, various host animals can be immunized by injection with the peptide corresponding to an REF1 epitope including but not limited to rabbits, mice, rats, sheep, goats, etc. In a preferred embodiment, the peptide is conjugated to an immunogenic carrier (*e.g.*, diphtheria toxoid, bovine serum albumin (BSA), or keyhole limpet hemocyanin (KLH)). Various adjuvants may be used to increase the immunological response, depending on the host species, including but not limited to Freund's (complete and incomplete), mineral gels (*e.g.*, aluminum hydroxide), surface active substances (*e.g.*, lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, keyhole limpet hemocyanins, dinitrophenol, and potentially useful human adjuvants such as BCG (Bacille Calmette-Guerin) and Corynebacterium parvum).

For preparation of monoclonal antibodies directed toward an REF1, it is contemplated that any technique that provides for the production of antibody molecules by continuous cell lines in culture finds use with the present invention (See *e.g.*, Harlow and Lane, Antibodies: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY). These include but are not limited to the hybridoma technique originally developed by Köhler and Milstein (Köhler and Milstein (1975) Nature, 256:495-497), as well as the trioma technique, the human B-cell hybridoma technique (See *e.g.*, Kozbor *et al.* (1983) Immunol. Tod., 4:72), and the EBV-hybridoma technique to produce human monoclonal antibodies (Cole *et al.* (1985) *in* Monoclonal Antibodies and Cancer Therapy, Alan R. Liss, Inc., pp. 77-96).

In addition, it is contemplated that techniques described for the production of single chain antibodies (U.S. Patent 4,946,778) find use in producing an REF1-specific single chain antibodies. An additional embodiment of the invention utilizes the techniques described for the construction of Fab expression libraries (Huse *et al.* (1989) Science, 246:1275-1281) to allow rapid and easy identification of monoclonal Fab fragments with the desired specificity for an REF1.

It is contemplated that any technique suitable for producing antibody fragments finds use in generating antibody fragments that contain the idiotype (antigen binding region) of the antibody molecule. For example, such fragments include but are not limited to: F(ab')2 fragment that can be produced by pepsin digestion of the antibody molecule; Fab' fragments that can be generated by reducing the disulfide bridges of the F(ab')2 fragment,

and Fab fragments that can be generated by treating the antibody molecule with papain and a reducing agent.

In the production of antibodies, it is contemplated that screening for the desired antibody is accomplished by techniques known in the art (*e.g.*, radioimmunoassay, ELISA (enzyme-linked immunosorbant assay), "sandwich" immunoassays, immunoradiometric assays, gel diffusion precipitin reactions, immunodiffusion assays, *in situ* immunoassays (*e.g.*, using colloidal gold, enzyme or radioisotope labels, for example), Western blots, precipitation reactions, agglutination assays (*e.g.*, gel agglutination assays, hemagglutination assays, etc.), complement fixation assays, immunofluorescence assays, protein A assays, and immunoelectrophoresis assays, etc.

In one embodiment, antibody binding is detected by detecting a label on the primary antibody. In another embodiment, the primary antibody is detected by detecting binding of a secondary antibody or reagent to the primary antibody. In a further embodiment, the secondary antibody is labeled. Many methods are known in the art for detecting binding in an immunoassay and are within the scope of the present invention. As is well known in the art, the immunogenic peptide should be provided free of the carrier molecule used in any immunization protocol. For example, if the peptide was conjugated to KLH, it may be conjugated to BSA, or used directly, in a screening assay.

In some embodiments of the present invention, the foregoing antibodies are used in methods known in the art relating to the expression of an REG1 (*e.g.*, for Western blotting), measuring levels thereof in appropriate biological samples, etc. The antibodies can be used to detect REF1 in a biological sample from a plant. The biological sample can be an extract of a tissue, or a sample fixed for microscopic examination.

The biological samples are then be tested directly for the presence of REF1 using an appropriate strategy (*e.g.*, ELISA or radioimmunoassay) and format (*e.g.*, microwells, dipstick (*e.g.*, as described in International Patent Publication WO 93/03367), etc. Alternatively, proteins in the sample can be size separated (*e.g.*, by polyacrylamide gel electrophoresis (PAGE), in the presence or not of sodium dodecyl sulfate (SDS), and the presence of REF1 detected by immunoblotting (Western blotting). Immunoblotting techniques are generally more effective with antibodies generated against a peptide corresponding to an epitope of a protein, and hence, are particularly suited to the present invention.

III. REF1 Coding Sequences

The present invention provides compositions comprising purified nucleic acid sequences encoding any of the REF1 polypeptides described above or below. Coding sequences include but are not limited to genes, cDNA, and RNA.

Thus, the present invention provides compositions comprising purified nucleic acid sequences encoding an REF1, as well as nucleic acid sequences encoding variants of REF1, including homologs, mutants, or fragments, or fusion proteins thereof, as described above and below. In yet other embodiments, the nucleic acid sequences encode a portion of an REF1 which retains some functional characteristic of a REF1. Examples of functional characteristics include the ability to act as an immunogen to produce an antibody which recognizes a REF1.

Coding sequences for REF1 include sequences isolated from an organism, which either comprises the coding sequence naturally or is transgenic and comprises a heterologous REF1 coding sequence, sequences which are chemically synthesized, as well as sequences which represent a combination of isolated and synthesized (as, for example, where isolated sequences are mutagenized, or where a sequence comprises parts of sequences isolated from different sources and/or synthesized from different sources).

Thus, in some embodiments of the invention, the coding sequence of a REF1 is synthesized, whole or in part, using chemical methods well known in the art (See *e.g.*, Caruthers *et al.* (1980) Nucl. Acids Res. Symp. Ser. 7:215-233; Crea and Horn (1980) Nucl. Acids Res. 9:2331; Matteucci and Caruthers (1980) Tetrahedron Lett. 21:719; and Chow and Kempe (1981) Nucl. Acids Res. 9:2807-2817

A. Arabidopsis Coding Sequence

In some embodiments, the sequences encode an Arabidopsis REF1. In some embodiments, the sequences comprise the sequence shown in Fig. 4 Panels A and B (SEQ ID Nos:1 or 2); in other embodiments, the sequences encode the amino acid sequence shown in Fig. 4 Panel C (SEQ ID NO:3).

B. Variant REF1 Coding Sequences

In other embodiments, the sequences encode a variant of the disclosed REF1 polypeptides; these variants include mutants, fragments, fusion proteins or functional equivalents of REF1. Exemplary sequences encoding variants are described further below.

C. Additional REF1 Coding Sequences and Genes

The present invention provides isolated nucleic acid sequences encoding REF1 in addition to those described above. For example, some embodiments of the present invention provide isolated polynucleotide sequences that are capable of hybridizing to SEQ ID NOs: 1 or 2 under conditions of low to high stringency as long as the polynucleotide sequence capable of hybridizing encodes a protein that retains a desired biological activity of RFEF1 as described above. In preferred embodiments, hybridization conditions are based on the melting temperature (T_m) of the nucleic acid binding complex and confer a defined "stringency" as explained above (See *e.g.*, Wahl *et al.* (1987) Meth. Enzymol., 152:399-407, incorporated herein by reference).

In other embodiments, an isolated nucleic acid sequence encoding an REF1 which is homologous to an Arabidopsis REF1 is provided; Exemplary but non-limiting examples are shown in Table 2.

In other embodiments of the present invention, alleles of an REF1 are provided. In preferred embodiments, alleles result from a mutation, (*i.e.*, a change in the nucleic acid sequence) and generally produce altered mRNAs or polypeptides whose structure or function may or may not be altered. Any given gene may have none, one or many allelic forms. Common mutational changes that give rise to alleles are generally ascribed to deletions, additions or substitutions of nucleic acids. Each of these types of changes may occur alone, or in combination with the others, and at the rate of one or more times in a given sequence.

These additional REF1 genes are discovered by the methods such as are described below.

IV. Methods of Identifying REF1 Coding Sequences and Genes

Other embodiments of the present invention provide methods to isolate nucleic acid sequences encoding REF1.

In some embodiments, a method involves obtaining a non-cDNA library for REF1 by using RT-PCR with degenerated primers to give a partial length clone, and subsequently using 3' and 5' RACE to define the 3' and 5' cDNA ends. A full length cDNA clone is then obtained via RT-PCR using primers based on the sequence of the 3' and 5' RACE products; this clone is then used to confirm the identity of the encoded polypeptide as an REF1. Confirmation of the identity of the encoded polypeptide includes expressing the polypeptide of the sequence encoding a putative REF1 (egg. the full length cDNA clone), and

characterizing the polypeptide of the putative REF1 coding sequence. Characterization includes but is not limited to detecting the presence of the expressed polypeptide by antibody-binding (REF1) or by detecting the reaction products of the expressed polypeptide as in any of the REF1 assays described above.

In some other embodiments, methods involve the preparation of a cDNA library from tissue. The cDNA library may be screened by hybridization with a REF1 probe (obtained, for example, from SEQ ID NOs:1 or 2). cDNA clones are identified which appear to encode a REF1; in other embodiments, cDNA clones are identified which appear to code for a portion of a REF1, which can be assembled into or utilized to create a complete coding sequence. Further embodiments include confirmation of a coding sequence as an REF1, as described above.

In yet other embodiments, methods involve first an examination of a plant expressed sequence tag (EST) database, in order to discover novel potential REF1 encoding sequences. In some embodiments, examination of a plant EST database involves blasting the database with the amino acid sequence of Arabidopsis REF1 (for example, SEQ ID NO:3), in order to discover ESTs encoding amino acid sequences with homology to Arabidopsis REF1 protein. In some further embodiments, the methods involve next assembling a clone encoding a complete putative REF1, and characterizing the expression products of such sequences so discovered as described above. In other further embodiments, these methods next involve sequencing likely candidate sequences, and characterizing the expression products of such sequences so discovered as described above.

In some embodiments, REF1 coding sequences, discovered by the methods of the present invention, can also be used to identify and isolate other plant genes. To isolate a gene, a ^{32}P-radiolabeled REF1 coding sequence (or cDNA) is used to screen, by DNA-DNA hybridization, a genomic library constructed from a plant genomic DNA. Single isolated clones that test positive for hybridization are proposed to contain part or all of an REF1 gene, and are sequenced. The sequence of these positive cloned plant genomic DNA is used to confirm the identity of the gene as an REF1. If a particular clone encodes only part of the gene, additional clones that test positive for hybridization to the REF1 coding sequence (or cDNA) are isolated and sequenced. Comparison of the full-length sequence of a putative REF1 gene to a cDNA are used to determine the location of introns, if they are present.

In other embodiments of the present invention, upstream sequences such as promoters and regulatory elements of a gene encoding an REF1 are detected by extending the gene by utilizing a nucleotide sequence encoding REF1 (*e.g.*, comprising SEQ ID Nos:1

or 2) in various methods known in the art. In some embodiments, it is contemplated that polymerase chain reaction (PCR) finds use in the present invention. This is a direct method that uses universal primers to retrieve unknown sequence adjacent to a known locus (Gobinda *et al.* (1993) PCR Methods Applic., 2:318-322). First, genomic DNA is amplified in the presence of primer to a linker sequence and a primer specific to the known region. The amplified sequences are then subjected to a second round of PCR with the same linker primer and another specific primer internal to the first one. Products of each round of PCR are transcribed with an appropriate RNA polymerase and sequenced using reverse transcriptase.

In another embodiments, inverse PCR is used to amplify or extend sequences using divergent primers based on a known region (Triglia *et al.* (1988) Nucleic Acids Res., 16:8186). In yet other embodiments of the present invention, capture PCR (Lagerstrom *et al.* (1991) PCR Methods Applic., 1:111-119) is used. In still other embodiments, walking PCR is utilized. In yet other embodiments of the present invention, add TAIL PCR is used as a preferred method for obtaining flanking genomic regions, including regulatory regions (Lui and Whittier, (1995); Lui *et al.* (1995)).

Preferred libraries for screening for full-length cDNAs include libraries that have been size-selected to include larger cDNAs. Also, random primed libraries are preferred, in that they contain more sequences that contain the 5' and upstream gene regions. A randomly primed library may be particularly useful in cases where an oligo d(T) library does not yield full-length cDNA. Genomic libraries are useful for obtaining introns and extending 5' sequence.

It is contemplated that the methods described above are used to discover other REF1 coding sequences and genes from plants. Exemplary plants are listed in Table 1.

V. Variants of REF1

In some embodiments, the present invention provides isolated variants of the disclosed nucleic acid sequence encoding REF1, and the polypeptides encoded thereby; these variants include mutants, fragments, fusion proteins, or functional equivalents of REF1. Thus, nucleotide sequences of the present invention are engineered in order to alter an REF1 coding sequence for a variety of reasons, including but not limited to alterations that modify the cloning, processing and/or expression of the gene product (such alterations include inserting new restriction sites, altering glycosylation patterns, and changing codon preference) as well as varying the enzymatic activity (such changes include but are not

limited to differing substrate affinities, differing substrate preferences and utilization, differing inhibitor affinities or effectiveness, differing reaction kinetics, varying subcellular localization, and varying protein processing and/or stability). For example, mutations are introduced which alter the substrate specificity, such that the preferred substrate is changed.

A. Mutants and Homologs of a Plant REF1

Some embodiments of the present invention provide mutant forms of an REF1 (*i.e.*, muteins). In preferred embodiments, variants result from mutation, (*i.e.*, a change in the nucleic acid sequence) and generally produce altered mRNAs or polypeptides whose structure or function may or may not be altered. Any given gene may have none, one, or many mutant forms. Common mutational changes that give rise to variants are generally ascribed to deletions, additions or substitutions of nucleic acids. Each of these types of changes may occur alone, or in combination with the others, and at the rate of one or more times in a given sequence.

Still other embodiments of the present invention provide isolated nucleic acid sequence encoding REF1 homologs, and the polypeptides encoded thereby.

It is contemplated that is possible to modify the structure of a peptide having an activity (*e.g.*, REF1 activity) for such purposes as increasing synthetic activity or altering the affinity of the REF1 for a substrate, or for increasing stability or turnover or subcellular location of the polypeptide. Such modified peptides are considered functional equivalents of peptides having an activity of an REF1 as defined herein. A modified peptide can be produced in which the nucleotide sequence encoding the polypeptide has been altered, such as by substitution, deletion, or addition.

In some preferred embodiments of the present invention, the alteration increases synthetic activity or alters the affinity of the REF1 for its substrate. In particularly preferred embodiments, these modifications do not significantly reduce the synthetic activity of the modified enzyme. In other words, construct "X" can be evaluated in order to determine whether it is a member of the genus of modified or variant REF1 of the present invention as defined functionally, rather than structurally. In preferred embodiments, the activity of variant REF1 is evaluated by the methods described above and in the Examples. Accordingly, in some embodiments the present invention provides nucleic acids encoding an REF1 that complement the coding region of SEQ ID NOs: 1 or 2.

In other preferred embodiments of the alteration, the alteration results in intracellular half-lives dramatically different from that of the corresponding wild-type protein. For

example, an altered protein is rendered either more stable or less stable to proteolytic degradation or other cellular process that result in destruction of, or otherwise inactivate REF1. Such homologs, and the genes that encode them, can be utilized to alter the activity of REF1 by modulating the half-life of the protein. For instance, a short half-life can give rise to more transient REF1 biological effects. Other variants have characteristics which are either similar to wild-type REF1, or which differ in one or more respects from wild-type REF1.

As described above, mutant forms of an REF1 are also contemplated as being equivalent to those peptides and DNA molecules that are set forth in more detail herein. For example, it is contemplated that isolated replacement of a leucine with an isoleucine or valine, an aspartate with a glutamate, a threonine with a serine, or a similar replacement of an amino acid with a structurally related amino acid (*i.e.*, conservative mutations) will not have a major effect on the biological activity of the resulting molecule. Accordingly, some embodiments of the present invention provide variants of an REF1 disclosed herein containing conservative replacements. Conservative replacements are those that take place within a family of amino acids that are related in their side chains. Genetically encoded amino acids can be divided into four families: (1) acidic (aspartate, glutamate); (2) basic (lysine, arginine, histidine); (3) nonpolar (alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine, tryptophan); and (4) uncharged polar (glycine, asparagine, glutamine, cysteine, serine, threonine, tyrosine). Phenylalanine, tryptophan, and tyrosine are sometimes classified jointly as aromatic amino acids. In similar fashion, the amino acid repertoire can be grouped as (1) acidic (aspartate, glutamate); (2) basic (lysine, arginine, histidine), (3) aliphatic (glycine, alanine, valine, leucine, isoleucine, serine, threonine), with serine and threonine optionally be grouped separately as aliphatic-hydroxyl; (4) aromatic (phenylalanine, tyrosine, tryptophan); (5) amide (asparagine, glutamine); and (6) sulfur - containing (cysteine and methionine) (*e.g.*, Stryer ed. (1981) *Biochemistry*, pg. 17-21, 2nd ed, WH Freeman and Co.). Whether a change in the amino acid sequence of a peptide results in a functional homolog can be readily determined by assessing the ability of the variant peptide to function in a fashion similar to the wild-type protein. Peptides having more than one replacement can readily be tested in the same manner.

More rarely, a variant includes "nonconservative" changes (*e.g.*, replacement of a glycine with a tryptophan). Analogous minor variations can also include amino acid deletions or insertions, or both. Guidance in determining which amino acid residues can be

substituted, inserted, or deleted without abolishing biological activity can be found using computer programs (*e.g.*, LASERGENE software, DNASTAR Inc., Madison, Wis.).

Mutants of an REF1 can be generated by any suitable method well known in the art, including but not limited to site-directed mutagenesis, randomized "point" mutagenesis, and domain-swap mutagenesis in which portions of the Arabidopsis REF1 cDNA are "swapped" with the analogous portion of other plant or bacterial REF1-encoding cDNAs (Back and Chappell (1996) PNAS 93: 6841-6845).

Variants may be produced by methods such as directed evolution or other techniques for producing combinatorial libraries of variants. Thus, the present invention further contemplates a method of generating sets of combinatorial mutants of the present REF1 proteins, as well as truncation mutants, and is especially useful for identifying potential variant sequences (*i.e.*, homologs) that possess the biological activity of an REF1 of the present invention. In addition, screening such combinatorial libraries is used to generate, for example, novel REF1 homologs that possess novel substrate specificities or other biological activities all together; examples of substrate specificities are described above.

It is contemplated that nucleic acids encoding REF1 (*e.g.*, SEQ ID NOs: 1 and 2, and fragments and variants thereof) can be utilized as starting nucleic acids for directed evolution. These techniques can be utilized to develop REF1 variants having desirable properties.

In some embodiments, artificial evolution is performed by random mutagenesis (*e.g.*, by utilizing error-prone PCR to introduce random mutations into a given coding sequence). This method requires that the frequency of mutation be finely tuned. As a general rule, beneficial mutations are rare, while deleterious mutations are common. This is because the combination of a deleterious mutation and a beneficial mutation often results in an inactive enzyme. The ideal number of base substitutions for targeted gene is usually between 1.5 and 5 (Moore and Arnold (1996) Nat. Biotech., 14, 458-67; Leung *et al.* (1989) Technique, 1:11-15; Eckert and Kunkel (1991) PCR Methods Appl., 1:17-24; Caldwell and Joyce (1992) PCR Methods Appl., 2:28-33; and Zhao and Arnold (1997) Nuc. Acids. Res., 25:1307-08). After mutagenesis, the resulting clones are selected for desirable activity. Successive rounds of mutagenesis and selection are often necessary to develop enzymes with desirable properties. It should be noted that only the useful mutations are carried over to the next round of mutagenesis.

In other embodiments of the present invention, the polynucleotides of the present invention are used in gene shuffling or sexual PCR procedures (*e.g.*, Smith (1994) Nature,

370:324-25; U.S. Pat. Nos. 5,837,458; 5,830,721; 5,811,238; 5,733,731). Gene shuffling involves random fragmentation of several mutant DNAs followed by their reassembly by PCR into full length molecules. Examples of various gene shuffling procedures include, but are not limited to, assembly following DNase treatment, the staggered extension process (STEP), and random priming *in vitro* recombination. In the DNase mediated method, DNA segments isolated from a pool of positive mutants are cleaved into random fragments with DNaseI and subjected to multiple rounds of PCR with no added primer. The lengths of random fragments approach that of the uncleaved segment as the PCR cycles proceed, resulting in mutations in present in different clones becoming mixed and accumulating in some of the resulting sequences. Multiple cycles of selection and shuffling have led to the functional enhancement of several enzymes (Stemmer (1994) Nature, 370:398-91; Stemmer (1994) Proc. Natl. Acad. Sci. USA, 91, 10747-10751; Crameri *et al.* (1996) Nat. Biotech., 14:315-319; Zhang *et al.* (1997) Proc. Natl. Acad. Sci. USA, 94:4504-09; and Crameri *et al.* (1997) Nat. Biotech., 15:436-38). Variants produced by directed evolution can be screened for REF1 activity by the methods described above and in the Examples.

In some embodiments of a combinatorial mutagenesis approach of the present invention, the amino acid sequences of a population of REF1 coding sequences are aligned, preferably to promote the highest homology possible. Such a population of variants can include, for example, REF1 homologs from one or more species, or REF1 homologs from the same species but which differ due to mutation. Amino acids that appear at each position of the aligned sequences are selected to create a degenerate set of combinatorial sequences.

In preferred embodiments of the present invention, the combinatorial REF1 library is produced by way of a degenerate library of genes encoding a library of polypeptides that each include at least a portion of candidate REF1-protein sequences. For example, a mixture of synthetic oligonucleotides is enzymatically ligated into gene sequences such that the degenerate set of candidate REF1 sequences are expressible as individual polypeptides, or alternatively, as a set of larger fusion proteins (*e.g.*, for phage display) containing the set of REF1 sequences therein.

There are many ways by which the library of potential REF1 homologs can be generated from a degenerate oligonucleotide sequence. In some embodiments, chemical synthesis of a degenerate gene sequence is carried out in an automatic DNA synthesizer, and the synthetic genes are ligated into an appropriate gene for expression. The purpose of a degenerate set of genes is to provide, in one mixture, all of the sequences encoding the desired set of potential REF1 sequences. The synthesis of degenerate oligonucleotides is

well known in the art (See *e.g.*, Narang (1983) Tetrahedron Lett., 39:3-9; Itakura *et al.* (1981) Recombinant DNA, in Walton (ed.), Proceedings of the 3rd Cleveland Symposium on Macromolecules, Elsevier, Amsterdam, pp 273-289; Itakura *et al.* (1984) Annu. Rev. Biochem., 53:323; Itakura *et al.* (1984) Science 198:1056; Ike *et al.* (1983) Nucl. Acid Res., 11:477). Such techniques have been employed in the directed evolution of other proteins (See *e.g.*, Scott *et al.* (1980) Science, 249:386-390; Roberts *et al.* (1992) Proc. Natl. Acad. Sci. USA, 89:2429-2433; Devlin *et al.* (1990) Science, 249: 404-406; Cwirla *et al.* (1990) Proc. Natl. Acad. Sci. USA, 87: 6378-6382; as well as U.S. Pat. Nos. 5,223,409, 5,198,346, and 5,096,815).

B. Truncation Mutants of Plant REF1

In addition, the present invention provides isolated nucleic acid sequences encoding fragments of REF1 (*i.e.*, truncation mutants), and the polypeptides encoded by such nucleic acid sequences. In preferred embodiments, an REF1 fragment is biologically active.

In some embodiments of the present invention, when expression of a portion of an REF1 protein is desired, it may be necessary to add a start codon (ATG) to the oligonucleotide fragment containing the desired sequence to be expressed. It is well known in the art that a methionine at the N-terminal position can be enzymatically cleaved by the use of the enzyme methionine aminopeptidase (MAP). MAP has been cloned from *E. coli* (Ben-Bassat *et al.* (1987) J. Bacteriol., 169:751-757) and Salmonella typhimurium and its *in vitro* activity has been demonstrated on recombinant proteins (Miller *et al.* (1990) Proc. Natl. Acad. Sci. USA, 84:2718-1722). Therefore, removal of an N-terminal methionine, if desired, can be achieved either *in vivo* by expressing such recombinant polypeptides in a host that produces MAP (*e.g.*, *E. coli* or CM89 or S. cerevisiae), or *in vitro* by use of purified MAP.

C. Fusion Proteins Containing Plant REF1

The present invention also provides nucleic acid sequences encoding fusion proteins incorporating all or part of REF1, and the polypeptides encoded by such nucleic acid sequences. In some embodiments, the fusion proteins have an REF1 functional domain with a fusion partner. Accordingly, in some embodiments of the present invention, the coding sequences for the polypeptide (*e.g.*, an REF1 functional domain) is incorporated as a part of a fusion gene including a nucleotide sequence encoding a different polypeptide. In one embodiment, a single fusion product polypeptide has SALDH or CALDH activity).

In some embodiments of the present invention, chimeric constructs code for fusion proteins containing a portion of an REF1 and a portion of another gene. In some embodiments, the fusion proteins have biological activity similar to the wild type EF1 (*e.g.*, have at least one desired biological activity of REF1). In other embodiments, the fusion proteins have altered biological activity.

In other embodiments of the present invention, chimeric constructs code for fusion proteins containing an REF1 gene or portion thereof and a leader or other signal sequences which direct the protein to targeted subcellular locations. Such sequences are well known in the art, and direct proteins to locations such as the chloroplast, the mitochondria, the endoplasmic reticulum, the tonoplast, the golgi network, and the plasmalemma.

In addition to utilizing fusion proteins to alter biological activity, it is widely appreciated that fusion proteins can also facilitate the expression and/or purification of proteins, such as an REF1 protein of the present invention. Accordingly, in some embodiments of the present invention, an REF1 is generated as a glutathione-S-transferase (*i.e.*, GST fusion protein). It is contemplated that such GST fusion proteins enables easy purification of an REF1, such as by the use of glutathione-derivatized matrices (*See e.g.*, Ausabel *et al.* (eds.) (1991) Current Protocols in Molecular Biology, John Wiley & Sons, NY).

In another embodiment of the present invention, a fusion gene coding for a purification leader sequence, such as a poly-(His)/enterokinase cleavage site sequence at the N-terminus of the desired portion of an REF1 allows purification of the expressed REF1 fusion protein by affinity chromatography using a Ni^{2+} metal resin. In still another embodiment of the present invention, the purification leader sequence is then subsequently removed by treatment with enterokinase (*See e.g.*, Hochuli *et al.* (1987) J. Chromatogr., 411:177; and Janknecht *et al.* Proc. Natl. Acad. Sci. USA, 88:8972). In yet other embodiments of the present invention, a fusion gene coding for a purification sequence appended to either the N (amino) or the C (carboxy) terminus allows for affinity purification; one example is addition of a hexahistidine tag to the carboxy terminus of an REF1, which is contemplated to be useful for affinity purification.

Techniques for making fusion genes are well known. Essentially, the joining of various nucleic acid fragments coding for different polypeptide sequences is performed in accordance with conventional techniques, employing blunt-ended or stagger-ended termini for ligation, restriction enzyme digestion to provide for appropriate termini, filling-in of cohesive ends as appropriate, alkaline phosphatase treatment to avoid undesirable joining,

and enzymatic ligation. In another embodiment of the present invention, the fusion gene can be synthesized by conventional techniques including automated DNA synthesizers. Alternatively, in other embodiments of the present invention, PCR amplification of gene fragments is carried out using anchor primers that give rise to complementary overhangs between two consecutive gene fragments that can subsequently be annealed to generate a chimeric gene sequence (*See e.g.*, Current Protocols in Molecular Biology, supra).

D. Screening Gene Products

A wide range of techniques are known in the art for screening gene products of combinatorial libraries made by point mutations, and for screening cDNA libraries for gene products having a certain property. Such techniques are generally adaptable for rapid screening of the gene libraries generated by the combinatorial mutagenesis of REF1 homologs. The most widely used techniques for screening large gene libraries typically comprise cloning the gene library into replicable expression vectors, transforming appropriate cells with the resulting library of vectors, and expressing the combinatorial genes under conditions in which detection of a desired activity facilitates relatively easy isolation of the vector encoding the gene whose product was detected. Techniques amenable to high through-put analysis as necessary to screen large numbers of degenerate sequences created by combinatorial mutagenesis techniques are also known.

In light of the present disclosure, other forms of mutagenesis generally applicable will be apparent to those skilled in the art in addition to the aforementioned rational mutagenesis based on conserved versus non-conserved residues. For example, REF1 homologs can be generated and screened using, for example, alanine scanning mutagenesis and the like (Ruf *et al.* (1994) Biochem., 33:1565-1572; Wang *et al.* (1994) J. Biol. Chem., 269:3095-3099; Balint (1993) Gene 137:109-118; Grodberg *et al.* (1993) Eur. J. Biochem., 218:597-601; Nagashima *et al.* (1993) J. Biol. Chem., 268:2888-2892; Lowman *et al.* (1991) Biochem., 30:10832-10838; and Cunningham *et al.* (1989) Science, 244:1081-1085), by linker scanning mutagenesis (Gustin *et al.* (1993) Virol., 193:653-660; Brown *et al.* (1992) Mol. Cell. Biol., 12:2644-2652; McKnight *et al.* Science, 232:316); or by saturation mutagenesis (Meyers *et al.* (1986) Science, 232:613).

VI. Utility of REF1

A. Commercial Utility of *REF1*

The present invention also provides methods of using *REF1* genes. Several uses of REF1 are contemplated, based upon the unexpected finding that REF1 oxidizes hydroxycinnamaldehydes to hydroxycinnamic acids. In some of these embodiments, the methods of the present invention are directed to decreasing hydroxycinnamic acid content in a plant cell. In some embodiments, the methods comprise growing a plant cell transfected with a construct directed to an REF1 gene, wherein the construct is able to decrease expression of REF1 in the cell, under conditions effective to decrease the hydroxycinnamic acid content of the plant cell. In other embodiments, the methods comprise transfecting a plant cell with a construct directed to an REF1 gene, wherein the construct is able to decrease expression of REF1 in the cell and growing the cell transfected with the construct under conditions effective to decrease the hydroxycinnamic acid content of the cell

For example, cell wall bound esters are principally ferulic acid esterified to cell wall pectic polysaccharides. This ferulic acid can be further cross linked to generate ester linked di-ferulic acid. In monocots, the hydroxycinnamic acid ferulate molecules plays a major role in cross-linking cell wall bound polysaccharides to monolignols (Grabber et al., 2000; Grabber et al., 2002). Further, kernel cell walls of major cereals, including but not limited to maize, wheat, rye, sorghum, triticale and Barley, contain considerable amounts of ferulic acid which impedes in cell wall digestibility when fed to monogastrics. *In vitro* enzyme digestion studies suggest that ferulate content and cell wall digestibility are negatively correlated (Grabber et al., 1998a,b). Identification of REF1 as an important enzyme responsible for ferulate synthesis provides a unique opportunity to alter the levels of ferulate content in cell walls and improve digestibility of cereal forage and kernel. Thus, in some embodiments, it is contemplated that the deposition of ferulic acid-dimers in plant cell walls can be modified, and preferably decreased, by altering the expression of *REF1*, and preferably decreasing it, thus increasing the digestibility of lignocellulosic materials used as animal feed. Decreasing ferulic acid content of plant cell walls is also contemplated to increase the ease of cellulose extraction in the paper industry.

Thus, in these embodiments, the present invention provides methods of decreasing the levels of ferulate in a plant cell, comprising decreasing the levels of *REF1* expression in the plant cell. Particularly suitable plants include forage crop plants, such as grasses and legumes, which are utilized for feed production and/or soil conservation. Non-limiting

examples include grasses annual brome, foxtail millet, slender wheat grass, Bermuda grass, brome grass, elephant grass (napier grass), fescue (meadow and tall, pearl millet, orchard grass (cocksfoot), pangola grass, redtop, red canary grass, perennial ryegrass, star grass, sudan grass, timothy, wheat grass (such as crested, fairway, and western), bahia grass, blue grass, buffel grass, dallis grass, guinea grass, kikuyu grass, weeping love grass, and buffalo grass, and legumes low hop clover, lespedeza (such as common and Korean), vetch (such as common and hairy), alfalfa (such as purple blossom, yellow blossom, and variegated), birdsfoot trefoil, alikse clove (such as alsike, red, and white) clover, and sweetclover (such white blossom and yellow blossom). Methods to decrease *REF1* expression are described in detail below.

In other examples, the hydroxycinnamic acid sinapine (also known as sinapoylcholine) is an anti-nutritional compound present in the seeds of commercially important oil seed, rapeseed (*Brassica napus*, of which canola, a high oleic acid variety, is an example), as well as in seeds of *Sinapis alba*, which is a distant relative of Brassica species and developed recently as an oil seed. The seed meal left after the oil is extracted from the harvested rapeseed and *S. alba* seed is typically utilized for animal feed. However, the presence of sinapine in this meal decreases its nutritional value as animal feed. Thus, in some embodiments, it is contemplated that the synthesis of sinapine during seed development can be modified, and preferably decreased, by altering the expression of *REF1*, and preferably decreasing it, thus increasing the nutritional value of the seed meal.

Thus, in these embodiments, the present invention provides methods of decreasing the levels of sinapine in a plant cell, comprising decreasing the levels of *REF1* expression in the plant cell. Particularly suitable plants include oil seed crop plants, such as members of the Brassica family. Methods to decrease *REF1* expression are described in detail below.

Other uses of *REF1* based upon the unexpected finding that REF1 oxidizes hydroxycinnamaldehydes to hydroxycinnamic acids generally. In these methods, the activity of REF1 is either increased or decreased in a plant cell, thereby modifying the hydroxycinnamic acid content of the plant cell. REF1 activity is increased by over-expressing REF1 coding sequences in a plant cell, as by transfecting a plant cell with additional copy of *REF1*, or with a variant of *REF1* with increased activity or increased stability, or by placing an REF1 coding sequence under control of a promoter which results in higher levels of expression of REF1. Methods for increasing *REF1* expression, and thus REF1 activity, are described in more detail below. REF1 activity is decreased as described above and below.

Other uses of REF1 are contemplated as a means to divert carbon within the phenylpropanoid pathway and thereby modify lignin quality and/or quantity. These uses are based upon the observation that REF1is an enzyme that oxidizes lignin precursors to their corresponding hydroxycinnamic acids. These products of coniferaldehyde are not involved in lignin biosynthesis (i.e., guaiacyl lignin and syringyl lignin) and are instead directed into soluble and cell wall-bound esters via ferulic acid and sinapic acid, both synthesized by REF1 from coniferaldehyde and sinapaldehyde, respectively (where sinapaldehyde is synthesized from coniferaldehyde via F5H and COMT) (see Figure 1).

Moreover, in Arabidopsis, the soluble esters include sinapoylmalate (in leaves), sinapoylcholine (in seeds), and sinapoylglucose, their common biosynthetic precursor. Soluble sinapic acid derivatives are relatively unusual in the plant kingdom, and are something of a hall mark of plants in the Brassicaceae such as Arabidopsis. However, these compounds derived from sinapic acid are not generally considered to be important lignin precursors. Thus, if carbon is diverted toward sinapic acid, most plants lack the enzyme needed to convert sinapic acid back to sinapoyl CoA, an obligatory step if the carbon is to be used in lignin biosynthesis. Thus, it is contemplated that *REF1* expression in lignifying tissues would decrease the syringyl units, sinapyl alcohol, available for lignin biosynthesis, and that total lignin content would be decreased as a result.

Thus, in some embodiments, it is contemplated that *REF1* over-expression, where *REF1* is expressed at low levels, or *REF1* expression, where *REF1* is not expressed, is used as a means to divert remove polymers from the guaiacyl lignin and syringyl lignin and into ferulic acid and sinapic acid esters. It is contemplated that over-expression of *REF1*, especially in tissue that do not express *REF1*, could divert the carbon flow from hydroxy-cinnamaldehydes (which form precursors to lignin synthesis) to hydroxy –cinnamic acids, thus reducing lignin content. Decreases in lignin content are generally associated with improved pulping characteristics and increased forage digestibility. It is further contemplated that in plants in which REF1 has a higher affinity for coniferaldehyde, over-expression of *REF1* results in coniferaldehyde specifically converted to ferulic acid, resulting in a specific reduction in guaiacyl lignin with out affecting the syringyl lignin. Thus, in these plants, it is contemplated that the amount and/or proportion of syringyl lignin increases, which results in improved pulping qualities of the plants as syringyl lignin is easily removed during the pulping process. Methods for increasing *REF1* expression, and thus REF1 activity, are described in more detail below.

B. Research Utility of REF1

In other embodiments, the *REF1* sequences are used for research purposes. For example, the discovery of the activity and role of REF1 can be used to more clearly define the biosynthesis of sinapate esters in Arabidopsis. Further, although sinapate ester synthesis is a hallmark of the Brassicaceae, sinapate conjugates are found in other species where in some cases they perform important functions. For example, the most abundant anthocyanin in wild carrot, *Daucus carota,* is a sinapoylated cyanidin glycoside (Harborne et al., 1983; Glä gen et al., 1992), and the enzyme catalyzing the "sinapoylation" of the non-acylated anthocyanin uses sinapoylglucose as the activated sinapate donor (Glä gen and Seitz, 1992). Since isolated *Daucus* vacuoles actively take up the sinapoylated anthocyanin, but not the non-acylated form (Hopp and Seitz, 1987), the sinapate moiety may function as a vacuolar uptake tag. Alternatively, the sinapate moiety may be required for, or be the site of, glutathione derivatization by a glutathione *S*-transferase analogous to that encoded by the maize *Bronze2* gene (Marrs et al., 1995).
It is contemplated that further analysis of the role of REF1 will result in a clearer understanding of sinapate ester synthesis and its regulation.

Moreover, in addition to the important role the phenylpropanoid pathway has in lignin biosynthesis, soluble phenylpropanoid compounds have a wide array of important functions in plants. They serve in the interaction of plants with their biotic and abiotic environments, mediate certain aspects of plant growth, development and pathogen resistance, and are important structural components of the plant secondary cell wall. For example, stilbenes and isoflavones are important phytoalexins in plants (Nicholson and Hammerschmidt, 1992). In maize and petunia, flavonoids have been shown to be necessary for pollen viability (Coe et al., 1981; Taylor and Jorgensen, 1992; van der Meer et al., 1992), and have been suggested to be endogenous modulators of auxin transport (Mathesius et al., 1998). Hydroxycinnamic acids lead to the synthesis of UV-sunscreens in plants (Landry et al., 1995), and are also precursors for lignin (Lewis and Yamomoto, 1990). Simpler phenylpropanoid-derived molecules such as acetosyringone act as signaling molecules in the interaction of plants with Agrobacterium (Stachel et al., 1985), and salicylate is a well-known mediator of plant defenses against pathogens (Dempsey et al., 1999). Lignan glycosides known as dihydrodiconiferyl glycosides (DCGs) have cytokinin-like activity in plants (Binns et al., 1987; Lynn et al., 1987; Teutonico et al., 1991; Orr and Lynn, 1992), and may be responsible for growth abnormalities seen in some transgenic plants in which phenylpropanoid metabolism has been perturbed (Tamagnone et al., 1998).

Phenylpropanoids are also increasingly being recognized as having an impact on human health. For example, isoflavones and lignans have beneficial estrogen-like activity in humans which is prompting their use as neutraceuticals (Humphreys and Chapple, 2000) and the stilbene resveratrol is thought to provide the health benefits associated with moderate wine consumption (Jang et al., 1997).

To understand the functions of phenylpropanoid metabolism in normal plant growth and development as well as in response to external biotic and abiotic stresses, it is essential to have a thorough knowledge of all of the catalysts involved. Thus, it is contemplated that *REF1* genes and variants thereof of the present invention are utilized in experiments designed to improve understanding of phenylpropanoid metabolism and its regulation.

VII. Methods of Modifying Plant Phenotype by Manipulating REF1 Activity in Plants

In some of the methods described above, it is contemplated that the nucleic acids encoding an REF1 of the present invention are utilized to either increase or decrease the level of REF1 expression (as mRNA and/or protein) in transfected cells as compared to the levels in wild-type cells.

A. Increased expression of REF1

Accordingly, in some embodiments, expression in plants of nucleic acid sequences encoding an REF1 of the present invention by the methods described above leads to the overexpression of REF1 in transgenic plants, plant tissues, or plant cells.

1. Transgenic Plants, Seeds, and Plant Parts

Plants are transformed with at least a heterologous gene encoding an REF1 of the present invention according to procedures well known in the art. It is contemplated that the heterologous gene is utilized to increase the level of the enzyme activities encoded by the heterologous gene.

a. Plants

The methods of the present invention are not limited to any particular plant. Indeed, a variety of plants are contemplated, including but not limited to tomato, potato, tobacco, pepper, rice, corn, barley, wheat, *Brassica*, *Arabidopsis*, sunflower, soybean, poplar, and pine. The group also includes non-agronomic species which are useful in developing

appropriate expression vectors such as tobacco, rapid cycling *Brassica* species, and *Arabidopsis thaliana*, and wild species undergoing domestication. In addition, plant lines where the endogenous *REF1* gene(s) has been inactivated by any method, but including mutagenesis (Katavic et al, 1995 and Zou et al. (1999), transposon tagging (Routaboul et al., 1999), and chimeraplasty may be utilized for expression of an *REF1* gene (for example, to confirm the identity of an isolated REF1 coding sequence, as further described in the Examples).

b. Vectors

The methods of the present invention contemplate the use of at least a heterologous gene encoding an REF1 of the present invention, as described above.

Heterologous genes intended for expression in plants are first assembled in expression cassettes comprising a promoter. Methods which are well known to those skilled in the art may be used to construct expression vectors containing a heterologous gene and appropriate transcriptional and translational control elements. These methods include *in vitro* recombinant DNA techniques, synthetic techniques, and *in vivo* genetic recombination. Such techniques are widely described in the art (See *e.g.*, Sambrook. *et al.* (1989) Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Press, Plainview, N.Y., and Ausubel, F. M. *et al.* (1989) Current Protocols in Molecular Biology, John Wiley & Sons, New York, N.Y).

In general, these vectors comprise a nucleic acid sequence of the invention encoding an REF1 of the present invention (as described above) operably linked to a promoter and other regulatory sequences (*e.g.*, enhancers, polyadenylation signals, etc.) required for expression in a plant.

Promoters include but are not limited to constitutive promoters, tissue-, organ-, and developmentally-specific promoters, and inducible promoters. Examples of promoters include but are not limited to: constitutive promoter 35S of cauliflower mosaic virus; a wound-inducible promoter from tomato, leucine amino peptidase ("LAP," Chao *et al.* (1999) Plant Physiol 120: 979-992); a chemically-inducible promoter from tobacco, Pathogenesis-Related 1 (PR1) (induced by salicylic acid and BTH (benzothiadiazole-7-carbothioic acid S-methyl ester)); a tomato proteinase inhibitor II promoter (PIN2) or LAP promoter (both inducible with methyl jasmonate); a heat shock promoter (US Pat 5,187,267); a tetracycline-inducible promoter (US Pat 5,057,422); and seed-specific promoters, such as those for seed storage proteins (*e.g.*, phaseolin, napin, oleosin, and a

promoter for soybean beta conglycin (Beachy *et al.* (1985) EMBO J. 4: 3047-3053)). All references cited herein are incorporated in their entirety.

The expression cassettes may further comprise any sequences required for expression of mRNA. Such sequences include, but are not limited to transcription terminators, enhancers such as introns, viral sequences, and sequences intended for the targeting of the gene product to specific organelles and cell compartments.

A variety of transcriptional terminators are available for use in expression of sequences using the promoters of the present invention. Transcriptional terminators are responsible for the termination of transcription beyond the transcript and its correct polyadenylation. Appropriate transcriptional terminators and those which are known to function in plants include, but are not limited to, the CaMV 35S terminator, the tml terminator, the pea rbcS E9 terminator, and the nopaline and octopine synthase terminator (See *e.g.*, Odell *et al.* (1985) Nature 313:810; Rosenberg *et al.* (1987) Gene, 56:125; Guerineau *et al.* (1991) Mol. Gen. Genet., 262:141; Proudfoot (1991) Cell, 64:671; Sanfacon *et al.* Genes Dev., 5:141 ; Mogen *et al.* (1990) Plant Cell, 2:1261; Munroe *et al.* (1990) Gene, 91:151; Ballad *et al.* (1989) Nucleic Acids Res. 17:7891; Joshi *et al.* (1987) Nucleic Acid Res., 15:9627).

In addition, in some embodiments, constructs for expression of the gene of interest include one or more of sequences found to enhance gene expression from within the transcriptional unit. These sequences can be used in conjunction with the nucleic acid sequence of interest to increase expression in plants. Various intron sequences have been shown to enhance expression, particularly in monocotyledonous cells. For example, the introns of the maize Adh1 gene have been found to significantly enhance the expression of the wild-type gene under its cognate promoter when introduced into maize cells (Calais *et al.* (1987) Genes Develop. 1: 1183). Intron sequences have been routinely incorporated into plant transformation vectors, typically within the non-translated leader.

In some embodiments of the present invention, the construct for expression of the nucleic acid sequence of interest also includes a regulator such as a nuclear localization signal (Calderone *et al.* (1984) Cell 39:499; Lassoer *et al.* (1991) Plant Molecular Biology 17:229), a plant translational consensus sequence (Joshi (1987) Nucleic Acids Research 15:6643), an intron (Luehrsen and Walbot (1991) Mol. Gen. Genet. 225:81), and the like, operably linked to the nucleic acid sequence encoding REF1.

In preparing a construct comprising a nucleic acid sequence encoding REF1 of the present invention, various DNA fragments can be manipulated, so as to provide for the

DNA sequences in the desired orientation (*e.g.*, sense or antisense) orientation and, as appropriate, in the desired reading frame. For example, adapters or linkers can be employed to join the DNA fragments or other manipulations can be used to provide for convenient restriction sites, removal of superfluous DNA, removal of restriction sites, or the like. For this purpose, *in vitro* mutagenesis, primer repair, restriction, annealing, resection, ligation, or the like is preferably employed, where insertions, deletions or substitutions (*e.g.*, transitions and transversions) are involved.

Numerous transformation vectors are available for plant transformation. The selection of a vector for use will depend upon the preferred transformation technique and the target species for transformation. For certain target species, different antibiotic or herbicide selection markers are preferred. Selection markers used routinely in transformation include the nptII gene which confers resistance to kanamycin and related antibiotics (Messing and Vierra (1982) Gene 19: 259; Bevan *et al.* (1983) Nature 304:184), the bar gene which confers resistance to the herbicide phosphinothricin (White *et al.* (1990) Nucl. Acids Res. 18:1062; Spencer *et al.* (1990) Theor. Appl. Genet. 79: 625), the hph gene which confers resistance to the antibiotic hygromycin (Blochlinger and Diggelmann (1984) Mol. Cell. Biol. 4:2929), and the dhfr gene, which confers resistance to methotrexate (Bourouis *et al.* (1983) EMBO J., 2:1099).

In some preferred embodiments, the vector is adapted for use in an *Agrobacterium* mediated transfection process (*See e.g.*, U.S. Pat. Nos. 5,981,839; 6,051,757; 5,981,840; 5,824,877; and 4,940,838; all of which are incorporated herein by reference). Construction of recombinant Ti and Ri plasmids in general follows methods typically used with the more common bacterial vectors, such as pBR322. Additional use can be made of accessory genetic elements sometimes found with the native plasmids and sometimes constructed from foreign sequences. These may include but are not limited to structural genes for antibiotic resistance as selection genes.

There are two systems of recombinant Ti and Ri plasmid vector systems now in use. The first system is called the "cointegrate" system. In this system, the shuttle vector containing the gene of interest is inserted by genetic recombination into a non-oncogenic Ti plasmid that contains both the cis-acting and trans-acting elements required for plant transformation as, for example, in the pMLJ1 shuttle vector and the non-oncogenic Ti plasmid pGV3850. The second system is called the "binary" system in which two plasmids are used; the gene of interest is inserted into a shuttle vector containing the cis-acting elements required for plant transformation. The other necessary functions are provided in

trans by the non-oncogenic Ti plasmid as exemplified by the pBIN19 shuttle vector and the non-oncogenic Ti plasmid PAL4404. Some of these vectors are commercially available.

In other embodiments of the invention, the nucleic acid sequence of interest is targeted to a particular locus on the plant genome. Site-directed integration of the nucleic acid sequence of interest into the plant cell genome may be achieved by, for example, homologous recombination using *Agrobacterium*-derived sequences. Generally, plant cells are incubated with a strain of *Agrobacterium* which contains a targeting vector in which sequences that are homologous to a DNA sequence inside the target locus are flanked by *Agrobacterium* transfer-DNA (T-DNA) sequences, as previously described (U.S. Pat. No. 5,501,967). One of skill in the art knows that homologous recombination may be achieved using targeting vectors which contain sequences that are homologous to any part of the targeted plant gene, whether belonging to the regulatory elements of the gene, or the coding regions of the gene. Homologous recombination may be achieved at any region of a plant gene so long as the nucleic acid sequence of regions flanking the site to be targeted is known.

In yet other embodiments, the nucleic acids of the present invention are utilized to construct vectors derived from plant (+) RNA viruses (*e.g.*, brome mosaic virus, tobacco mosaic virus, alfalfa mosaic virus, cucumber mosaic virus, tomato mosaic virus, and combinations and hybrids thereof). Generally, the inserted REF1 polynucleotide of the present invention can be expressed from these vectors as a fusion protein (*e.g.*, coat protein fusion protein) or from its own subgenomic promoter or other promoter. Methods for the construction and use of such viruses are described in U.S. Pat. Nos. 5,846,795; 5,500,360; 5,173,410; 5,965,794; 5,977,438; and 5,866,785, all of which are incorporated herein by reference.

In some embodiments of the present invention, where the nucleic acid sequence of interest is introduced directly into a plant. One vector useful for direct gene transfer techniques in combination with selection by the herbicide Basta (or phosphinothricin) is a modified version of the plasmid pCIB246, with a CaMV 35S promoter in operational fusion to the *E. coli* GUS gene and the CaMV 35S transcriptional terminator (WO 93/07278).

c. Transformation Techniques

Once a nucleic acid sequence encoding an REF1 of the present invention is operatively linked to an appropriate promoter and inserted into a suitable vector for the particular transformation technique utilized (*e.g.*, one of the vectors described above), the

recombinant DNA described above can be introduced into the plant cell in a number of art-recognized ways. Those skilled in the art will appreciate that the choice of method might depend on the type of plant targeted for transformation. In some embodiments, the vector is maintained episomally. In other embodiments, the vector is integrated into the genome.

In some embodiments, direct transformation in the plastid genome is used to introduce the vector into the plant cell (See *e.g.*, U.S. Patent Nos 5,451,513; 5,545,817; 5,545,818; PCT application WO 95/16783). The basic technique for chloroplast transformation involves introducing regions of cloned plastid DNA flanking a selectable marker together with the nucleic acid encoding the RNA sequences of interest into a suitable target tissue (*e.g.*, using biolistics or protoplast transformation with calcium chloride or PEG). The 1 to 1.5 kb flanking regions, termed targeting sequences, facilitate homologous recombination with the plastid genome and thus allow the replacement or modification of specific regions of the plastome. Initially, point mutations in the chloroplast 16S rRNA and rps12 genes conferring resistance to spectinomycin and/or streptomycin are utilized as selectable markers for transformation (Svab *et al.* (1990) PNAS, 87:8526; Staub and Maliga, (1992) Plant Cell, 4:39). The presence of cloning sites between these markers allowed creation of a plastid targeting vector introduction of foreign DNA molecules (Staub and Maliga (1993) EMBO J., 12:601). Substantial increases in transformation frequency are obtained by replacement of the recessive rRNA or r-protein antibiotic resistance genes with a dominant selectable marker, the bacterial aadA gene encoding the spectinomycin-detoxifying enzyme aminoglycoside-3'-adenyltransferase (Svab and Maliga (1993) PNAS, 90:913). Other selectable markers useful for plastid transformation are known in the art and encompassed within the scope of the present invention. Plants homoplasmic for plastid genomes containing the two nucleic acid sequences separated by a promoter of the present invention are obtained, and are preferentially capable of high expression of the RNAs encoded by the DNA molecule.

In other embodiments, vectors useful in the practice of the present invention are microinjected directly into plant cells by use of micropipettes to mechanically transfer the recombinant DNA (Crossway (1985) Mol. Gen. Genet, 202:179). In still other embodiments, the vector is transferred into the plant cell by using polyethylene glycol (Krens *et al.* (1982) Nature, 296:72; Crossway *et al.* (1986) BioTechniques, 4:320); fusion of protoplasts with other entities, either minicells, cells, lysosomes or other fusible lipid-surfaced bodies (Fraley *et al.* (1982) Proc. Natl. Acad. Sci., USA, 79:1859); protoplast

transformation (EP 0 292 435); direct gene transfer (Paszkowski *et al.* (1984) EMBO J., 3:2717; Hayashimoto *et al.* (1990) Plant Physiol. 93:857).

In still further embodiments, the vector may also be introduced into the plant cells by electroporation. (Fromm, *et al.* (1985) Proc. Natl Acad. Sci. USA 82:5824; Riggs *et al.* (1986) Proc. Natl. Acad. Sci. USA 83:5602). In this technique, plant protoplasts are electroporated in the presence of plasmids containing the gene construct. Electrical impulses of high field strength reversibly permeabilize biomembranes allowing the introduction of the plasmids. Electroporated plant protoplasts reform the cell wall, divide, and form plant callus.

In yet other embodiments, the vector is introduced through ballistic particle acceleration using devices (*e.g.*, available from Agracetus, Inc., Madison, Wis. and Dupont, Inc., Wilmington, Del). (See *e.g.*, U.S. Pat. No. 4,945,050; and McCabe *et al.* (1988) Biotechnology 6:923). See also, Weissinger *et al.* (1988) Annual Rev. Genet. 22:421; Sanford *et al.* (1987) Particulate Science and Technology, 5:27 (onion); Svab *et al.* (1990) Proc. Natl. Acad. Sci. USA, 87:8526 (tobacco chloroplast); Christou *et al.* (1988) Plant Physiol., 87:671 (soybean); McCabe *et al.* (1988) Bio/Technology 6:923 (soybean); Klein *et al.* (1988) Proc. Natl. Acad. Sci. USA, 85:4305 (maize); Klein *et al.* (1988) Bio/Technology, 6:559 (maize); Klein *et al.* (1988) Plant Physiol., 91:4404 (maize); Fromm *et al.* (1990) Bio/Technology, 8:833; and Gordon-Kamm *et al.* (1990) Plant Cell, 2:603 (maize); Koziel *et al.* (1993) Biotechnology, 11:194 (maize); Hill *et al.* (1995) Euphytica, 85:119 and Koziel *et al.* (1996) Annals of the New York Academy of Sciences 792:164; Shimamoto *et al.* (1989) Nature 338: 274 (rice); Christou *et al.* (1991) Biotechnology, 9:957 (rice); Datta *et al.* (1990) Bio/Technology 8:736 (rice); European Patent Application EP 0 332 581 (orchardgrass and other Pooideae); Vasil *et al.* (1993) Biotechnology, 11: 1553 (wheat); Weeks *et al.* (1993) Plant Physiol., 102: 1077 (wheat); Wan *et al.* (1994) Plant Physiol. 104: 37 (barley); Jahne *et al.* (1994) Theor. Appl. Genet. 89:525 (barley); Knudsen and Muller (1991) Planta, 185:330 (barley); Umbeck *et al.* (1987) Bio/Technology 5: 263 (cotton); Casas *et al.* (1993) Proc. Natl. Acad. Sci. USA 90:11212 (sorghum); Somers *et al.* (1992) Bio/Technology 10:1589 (oat); Torbert *et al.* (1995) Plant Cell Reports, 14:635 (oat); Weeks *et al.* (1993) Plant Physiol., 102:1077 (wheat); Chang *et al.*, WO 94/13822 (wheat) and Nehra *et al.* (1994) The Plant Journal, 5:285 (wheat).

In addition to direct transformation, in some embodiments, the vectors comprising a nucleic acid sequence encoding an REF1 of the present invention are transferred using Agrobacterium-mediated transformation (Hinchee *et al.* (1988) Biotechnology, 6:915;

Ishida *et al.* (1996) Nature Biotechnology 14:745). Agrobacterium is a representative genus of the gram-negative family Rhizobiaceae. Its species are responsible for plant tumors such as crown gall and hairy root disease. In the dedifferentiated tissue characteristic of the tumors, amino acid derivatives known as opines are produced and catabolized. The bacterial genes responsible for expression of opines are a convenient source of control elements for chimeric expression cassettes. Heterologous genetic sequences (*e.g.*, nucleic acid sequences of the present invention operatively linked to a promoter), can be introduced into appropriate plant cells, by means of the Ti plasmid of Agrobacterium tumefaciens. The Ti plasmid is transmitted to plant cells on infection by Agrobacterium tumefaciens, and is stably integrated into the plant genome (Schell (1987) Science, 237: 1176). Species which are susceptible infection by Agrobacterium may be transformed *in vitro*. Alternatively, plants may be transformed *in vivo*, such as by transformation of a whole plant by Agrobacteria infiltration of adult plants, as in a "floral dip" method (Bechtold N, Ellis J, Pelletier G (1993) Cr. Acad. Sci. III - Vie 316: 1194-1199).

d. Regeneration

After selecting for transformed plant material that can express the heterologous gene encoding an REF1 of the present invention, whole plants are regenerated. Plant regeneration from cultured protoplasts is described in Evans *et al.* (1983) Handbook of Plant Cell Cultures, Vol. 1: (MacMillan Publishing Co. New York); and Vasil I. R. (ed.), Cell Culture and Somatic Cell Genetics of Plants, Acad. Press, Orlando, Vol. I (1984), and Vol. III (1986). It is known that many plants can be regenerated from cultured cells or tissues, including but not limited to all major species of sugarcane, sugar beet, cotton, fruit and other trees, legumes and vegetables, and monocots (*e.g.*, the plants described above). Means for regeneration vary from species to species of plants, but generally a suspension of transformed protoplasts containing copies of the heterologous gene is first provided. Callus tissue is formed and shoots may be induced from callus and subsequently rooted.

Alternatively, embryo formation can be induced from the protoplast suspension. These embryos germinate and form mature plants. The culture media will generally contain various amino acids and hormones, such as auxin and cytokinins. Shoots and roots normally develop simultaneously. Efficient regeneration will depend on the medium, on the genotype, and on the history of the culture. The reproducibility of regeneration depends on the control of these variables.

e. Generation of Transgenic lines

Transgenic lines are established from transgenic plants by tissue culture propagation. The presence of nucleic acid sequences encoding a heterologous REF1 of the present invention (including mutants or variants thereof) may be transferred to related varieties by traditional plant breeding techniques.

These transgenic lines are then utilized for evaluation of lignin characteristics (such as quantity and/or quality), for hydroxycinnamic acid characteristics (such as quantity and/or quality), for altered digestibility, altered profiles of secondary metabolites, and/or for other agronomic traits.

B. Decreased Expression of REF1

In other embodiments of the present invention, the REF1 polynucleotides are utilized to decrease the level of REF1 protein or mRNA in transgenic plants, plant tissues, or plant cells as compared to wild-type plants, plant tissues, or plant cells. In these embodiments, the function of an *REF1* gene is disrupted by any effective technique, including but not limited to antisense, co-suppression, and RNA interference, as is described above and below.

As described above, in some embodiments, it is contemplated that the nucleic acids encoding an *REF1* polypeptide of the present invention may be utilized to decrease the level of REF1 mRNA and/or protein in transfected cells as compared to the levels in wild-type cells. In some of these embodiments, the nucleic acid sequence encoding an REF1 protein of the present invention is used to design a nucleic acid sequence encoding a nucleic acid product which interferes with the expression of the nucleic acid encoding a REF1 polypeptide, where the interference is based upon a coding sequence of the encoded REF1 polypeptide. Exemplary methods of construct are described further below. These constructs are then used to transfect plants as described above under the description of methods to increase *REF1* expression.

One method of reducing REF1 expression utilizes expression of antisense transcripts. Antisense RNA has been used to inhibit plant target genes in a tissue-specific manner (*e.g.*, van der Krol *et al.* (1988) Biotechniques 6:958-976). Antisense inhibition has been shown using the entire cDNA sequence as well as a partial cDNA sequence (*e.g.*, Sheehy *et al.* (1988) Proc. Natl. Acad. Sci. USA 85:8805-8809; Cannon *et al.* (1990) Plant Mol. Biol. 15:39-47). There is also evidence that 3' non-coding sequence fragment and 5' coding sequence fragments, containing as few as 41 base-pairs of a 1.87 kb cDNA, can play

important roles in antisense inhibition (Ch'ng *et al.* (1989) Proc. Natl. Acad. Sci. USA 86:10006-10010).

Accordingly, in some embodiments, an REF1 encoding-nucleic acid of the present invention (*e.g.*, SEQ ID NOs: 1 and 2, and fragments and variants thereof) are oriented in a vector and expressed so as to produce antisense transcripts. To accomplish this, a nucleic acid segment from the desired gene is cloned and operably linked to a promoter such that the antisense strand of RNA will be transcribed. The expression cassette is then transformed into plants and the antisense strand of RNA is produced. The nucleic acid segment to be introduced generally will be substantially identical to at least a portion of the endogenous gene or genes to be repressed. The sequence, however, need not be perfectly identical to inhibit expression. The vectors of the present invention can be designed such that the inhibitory effect applies to other proteins within a family of genes exhibiting homology or substantial homology to the target gene.

Furthermore, for antisense suppression, the introduced sequence also need not be full length relative to either the primary transcription product or fully processed mRNA. Generally, higher homology can be used to compensate for the use of a shorter sequence. Furthermore, the introduced sequence need not have the same intron or exon pattern, and homology of non-coding segments may be equally effective. Normally, a sequence of between about 30 or 40 nucleotides and about full length nucleotides should be used, though a sequence of at least about 100 nucleotides is preferred, a sequence of at least about 200 nucleotides is more preferred, and a sequence of at least about 500 nucleotides is especially preferred.

Catalytic RNA molecules or ribozymes can also be used to inhibit expression of the target gene or genes. It is possible to design ribozymes that specifically pair with virtually any target RNA and cleave the phosphodiester backbone at a specific location, thereby functionally inactivating the target RNA. In carrying out this cleavage, the ribozyme is not itself altered, and is thus capable of recycling and cleaving other molecules, making it a true enzyme. The inclusion of ribozyme sequences within antisense RNAs confers RNA-cleaving activity upon them, thereby increasing the activity of the constructs.

A number of classes of ribozymes have been identified. One class of ribozymes is derived from a number of small circular RNAs which are capable of self-cleavage and replication in plants. The RNAs replicate either alone (viroid RNAs) or with a helper virus (satellite RNAs). Examples include RNAs from avocado sunblotch viroid and the satellite RNAs from tobacco ringspot virus, lucerne transient streak virus, velvet tobacco mottle

virus, Solanum nodiflorum mottle virus and subterranean clover mottle virus. The design and use of target RNA-specific ribozymes is described in Haseloff, *et al.* (1988) Nature 334:585-591. Ribozymes targeted to the mRNA of a lipid biosynthetic gene, resulting in a heritable increase of the target enzyme substrate, have also been described (Merlo AO et al. (1998) Plant Cell 10: 1603-1621).

Another method of reducing REF1 expression utilizes the phenomenon of cosuppression or gene silencing (*See e.g.*, U.S. Pat. No. 6,063,947, incorporated herein by reference). The phenomenon of cosuppression has also been used to inhibit plant target genes in a tissue-specific manner. Cosuppression of an endogenous gene using a full-length cDNA sequence as well as a partial cDNA sequence (730 bp of a 1770 bp cDNA) are known (*e.g.*, Napoli *et al.* (1990) Plant Cell 2:279-289; van der Krol *et al.* (1990) Plant Cell 2:291-299; Smith *et al.* (1990) Mol. Gen. Genetics 224:477-481). Accordingly, in some embodiments the nucleic acid sequences encoding an REF1 of the present invention (*e.g.* including SEQ ID NOs 1 and 2, and fragments and variants thereof) are expressed in another species of plant to effect cosuppression of a homologous gene.

Generally, where inhibition of expression is desired, some transcription of the introduced sequence occurs. The effect may occur where the introduced sequence contains no coding sequence per se, but only intron or untranslated sequences homologous to sequences present in the primary transcript of the endogenous sequence. The introduced sequence generally will be substantially identical to the endogenous sequence intended to be repressed. This minimal identity will typically be greater than about 65%, but a higher identity might exert a more effective repression of expression of the endogenous sequences. Substantially greater identity of more than about 80% is preferred, though about 95% to absolute identity would be most preferred. As with antisense regulation, the effect should apply to any other proteins within a similar family of genes exhibiting homology or substantial homology.

For cosuppression, the introduced sequence in the expression cassette, needing less than absolute identity, also need not be full length, relative to either the primary transcription product or fully processed mRNA. This may be preferred to avoid concurrent production of some plants which are overexpressers. A higher identity in a shorter than full length sequence compensates for a longer, less identical sequence. Furthermore, the introduced sequence need not have the same intron or exon pattern, and identity of non-coding segments will be equally effective. Normally, a sequence of the size ranges noted above for antisense regulation is used.

An effective method to down regulate a gene is by hairpin RNA constructs. Guidance to the design of such constructs for efficient, effective and high throughput gene silencing have been described (Wesley SV *et al.* (2001) Plant J. 27: 581-590). Another method to decrease expression of a gene (either endogenous or exogenous) is via siRNAs. siRNAs can be applied to a plant and taken up by plant cells; alternatively, siRNAs can be expressed *in vivo* from an expression cassette.

An advantage of siRNAs is the short length of the mRNA which is targeted; this allows preferential targeting of a first sequence which is very similar to a second sequence, while allowing expression of the second, non-targeted sequence.

VIII. Expression of Cloned REF1

In other embodiment of the present invention, nucleic acid sequences corresponding to the *REF1* genes, homologs and mutants as described above may be used to generate recombinant DNA molecules that direct the expression of the encoded protein product in appropriate host cells.

As will be understood by those of skill in the art, it may be advantageous to produce REF1-encoding nucleotide sequences possessing non-naturally occurring codons. Therefore, in some preferred embodiments, codons preferred by a particular prokaryotic or eukaryotic host (Murray *et al.* (1989) Nucl. Acids Res., 17) can be selected, for example, to increase the rate of REF1 expression or to produce recombinant RNA transcripts having desirable properties, such as a longer half-life, than transcripts produced from naturally occurring sequence.

A. Vectors for Production of Plant REF1

The nucleic acid sequences of the present invention may be employed for producing polypeptides by recombinant techniques. Thus, for example, the nucleic acid sequence may be included in any one of a variety of expression vectors for expressing a polypeptide. In some embodiments of the present invention, vectors include, but are not limited to, chromosomal, nonchromosomal and synthetic DNA sequences (*e.g.*, derivatives of SV40, bacterial plasmids, phage DNA; baculovirus, yeast plasmids, vectors derived from combinations of plasmids and phage DNA, and viral DNA such as vaccinia, adenovirus, fowl pox virus, and pseudorabies). It is contemplated that any vector may be used as long as it is replicable and viable in the host.

In particular, some embodiments of the present invention provide recombinant constructs comprising one or more of the nucleic sequences as broadly described above (*e.g.*, SEQ ID NOs: 1 or 2). In some embodiments of the present invention, the constructs comprise a vector, such as a plasmid or viral vector, into which a nucleic acid sequence of the invention has been inserted, in a forward or reverse orientation. In preferred embodiments of the present invention, the appropriate nucleic acid sequence is inserted into the vector using any of a variety of procedures. In general, the nucleic acid sequence is inserted into an appropriate restriction endonuclease site(s) by procedures known in the art.

Large numbers of suitable vectors are known to those of skill in the art, and are commercially available. Such vectors include, but are not limited to, the following vectors: 1) Bacterial -- pQE70, pQE60, pQE-9 (Qiagen), pBS, pD10, phagescript, psiX174, pbluescript SK, pBSKS, pNH8A, pNH16a, pNH18A, pNH46A (Stratagene); ptrc99a, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia); and 2) Eukaryotic -- pWLNEO, pSV2CAT, pOG44, PXT1, pSG (Stratagene) pSVK3, pBPV, pMSG, and pSVL (Pharmacia). Any other plasmid or vector may be used as long as they are replicable and viable in the host. In some preferred embodiments of the present invention, plant expression vectors comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation sites, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences. In other embodiments, DNA sequences derived from the SV40 splice, and polyadenylation sites may be used to provide the required nontranscribed genetic elements.

In certain embodiments of the present invention, a nucleic acid sequence of the present invention within an expression vector is operatively linked to an appropriate expression control sequence(s) (promoter) to direct mRNA synthesis. Promoters useful in the present invention include, but are not limited to, the LTR or SV40 promoter, the *E. coli* lac or trp, the phage lambda P_L and P_R, T3 and T7 promoters, and the cytomegalovirus (CMV) immediate early, herpes simplex virus (HSV) thymidine kinase; and mouse metallothionein-I promoters and other promoters known to control expression of gene in prokaryotic or eukaryotic cells or their viruses. In other embodiments of the present invention, recombinant expression vectors include origins of replication and selectable markers permitting transformation of the host cell (*e.g.*, dihydrofolate reductase or neomycin resistance for eukaryotic cell culture, or tetracycline or ampicillin resistance in *E. coli*).

In some embodiments of the present invention, transcription of the DNA encoding polypeptides of the present invention by higher eukaryotes is increased by inserting an enhancer sequence into the vector. Enhancers are cis-acting elements of DNA, usually about from 10 to 300 bp that act on a promoter to increase its transcription. Enhancers useful in the present invention include, but are not limited to, the SV40 enhancer on the late side of the replication origin bp 100 to 270, a cytomegalovirus early promoter enhancer, the polyoma enhancer on the late side of the replication origin, and adenovirus enhancers.

In other embodiments, the expression vector also contains a ribosome binding site for translation initiation and a transcription terminator. In still other embodiments of the present invention, the vector may also include appropriate sequences for amplifying expression.

B. Host Cells for Production of Plant REF1

In a further embodiment, the present invention provides host cells containing any of the above-described constructs. In some embodiments of the present invention, the host cell is a higher eukaryotic cell (*e.g.*, a plant cell). In other embodiments of the present invention, the host cell is a lower eukaryotic cell (*e.g.*, a yeast cell). In still other embodiments of the present invention, the host cell can be a prokaryotic cell (*e.g.*, a bacterial cell). Specific examples of host cells include, but are not limited to, *Escherichia coli*, *Salmonella typhimurium*, *Bacillus subtilis*, and various species within the genera *Pseudomonas*, *Streptomyces*, and *Staphylococcus*, as well as *Saccharomycees cerivisiae*, *Schizosaccharomycees pombe*, *Drosophila* S2 cells, Spodoptera Sf9 cells, Chinese hamster ovary (CHO) cells, COS-7 lines of monkey kidney fibroblasts, (Gluzman (1981) Cell 23:175), 293T, C127, 3T3, HeLa and BHK cell lines, NT-1 (tobacco cell culture line), root cell and cultured roots in rhizosecretion (Gleba *et al.* (1999) Proc Natl Acad Sci USA 96: 5973-5977). Other examples include microspore-derived cultures of oilseed rape. (Weselake RJ and Taylor DC (1999) Prog. Lipid Res. 38: 401), and transformation of pollen and microspore culture systems. Further examples are described in the Examples.

The constructs in host cells can be used in a conventional manner to produce the gene product encoded by any of the recombinant sequences of the present invention described above. In some embodiments, introduction of the construct into the host cell can be accomplished by calcium phosphate transfection, DEAE-Dextran mediated transfection, or electroporation (See *e.g.*, Davis *et al.* (1986) Basic Methods in Molecular Biology).

Alternatively, in some embodiments of the present invention, a polypeptide of the invention can be synthetically produced by conventional peptide synthesizers.

Proteins can be expressed in eukaryotic cells, yeast, bacteria, or other cells under the control of appropriate promoters. Cell-free translation systems can also be employed to produce such proteins using RNAs derived from a DNA construct of the present invention. Appropriate cloning and expression vectors for use with prokaryotic and eukaryotic hosts are described by Sambrook, *et al.* (1989) Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor, N.Y.

In some embodiments of the present invention, following transformation of a suitable host strain and growth of the host strain to an appropriate cell density, the selected promoter is induced by appropriate means (*e.g.*, temperature shift or chemical induction) and cells are cultured for an additional period. In other embodiments of the present invention, cells are typically harvested by centrifugation, disrupted by physical or chemical means, and the resulting crude extract retained for further purification. In still other embodiments of the present invention, microbial cells employed in expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents.

IX. Production of Hydroxycinnamic Aldehydes

In yet other embodiments, the present invention provides methods to produce hydroxycinnamic acids by heterologous coding sequences of REF1 as described above.

A. *In vivo* in Transgenic Organism

In some embodiments of the present invention, hydroxycinnamic acids (HA) are produced *in vivo*, by providing an organism transformed with a heterologous gene encoding an REF1 of the present invention and growing the transgenic organism under conditions sufficient to effect production of HA. In other embodiments of the present invention, HA are produced *in vivo* by transforming an organism with a heterologous gene encoding an REF1 of the present invention and growing the transgenic organism under conditions sufficient to effect production of REF1. Illustrative examples of transgenic organisms are described below and provided in the Examples.

Organisms which are transformed with a heterologous gene encoding an REF1 of the present invention include preferably those which naturally synthesize and store in some manner aldehydes that are substrates for REF1 and those which are commercially feasible to

grow and suitable for harvesting large amounts of the HA products. Such organisms include but are not limited to, yeast and algae, and plants. Many commercial cultivars can be transformed with heterologous genes. In cases where that is not possible, non-commercial cultivars of plants can be transformed, and the trait for expression of REF1 of the present invention moved to commercial cultivars by breeding techniques well-known in the art.

A heterologous gene encoding an REF1 of the present invention, which includes variants of an REF1, includes any suitable sequence of the invention as described above. Preferably, the heterologous gene is provided within an expression vector such that transformation with the vector results in expression of the polypeptide; suitable vectors are described above and following.

A transgenic organism is grown under conditions sufficient to effect production of HA. In some embodiments of the present invention, a transgenic organism is supplied with exogenous substrates of the HA (as, for example, when the organism is grown in a fermenter). Such substrates can comprise any aldehyde that is a substrate for REF1, including but not limited to coniferaldehyde and sinapaldehyde. Substrates may be supplied in various forms as are well known in the art; such forms include aqueous suspensions prepared by sonication, aqueous suspensions prepared with detergents and other surfactants, dissolution of the substrate into a solvent, and dried powders of substrates. Such forms may be added to organisms or cultured cells or tissues grown in fermenters.

In yet other embodiments of the present invention, a transgenic organism comprises a heterologous gene encoding an REF1 of the present invention operably linked to an inducible promoter, and is grown either in the presence of the an inducing agent, or is grown and then exposed to an inducing agent. In still other embodiments of the present invention, a transgenic organism comprises a heterologous gene encoding an REF1 of the present invention operably linked to a promoter which is either tissue specific or developmentally specific, and is grown to the point at which the tissue is developed or the developmental stage at which the developmentally-specific promoter is activated. Such promoters include seed specific promoters.

In alternative embodiments, a transgenic organism as described above is engineered to produce greater amounts of REF1 aldehyde substrates, including but not limited to coniferaldehyde and sinapaldehyde.

In other embodiments of the present invention, a host organism produces large amounts of a desired substrate, including any aldehyde that is a substrate for REF1, including but not limited to coniferaldehyde and sinapaldehyde.

In other embodiments, a host organism produces low amounts of a desired substrate such as any aldehyde that is a substrate for REF1, including but not limited to coniferaldehyde and sinapaldehyde. Such hosts may occur naturally or via genetic engineering techniques.

In other embodiments of the present invention, the methods for producing HA further comprise collecting the HA produced. Such methods are known generally in the art, and include harvesting the transgenic organisms and extracting the HA, as for example is described in the Examples.

B. *In vitro* Systems

In other embodiments of the present invention, HA are produced *in vitro*, from either nucleic acid sequences encoding an REF1 of the present invention.

1. Using Nucleic Acid Sequences Encoding REF1

In some embodiments of the present invention, methods for producing HA comprise incubating an isolated nucleic acid sequence encoding an REF1 of the present invention in an *in vitro* expression systems under conditions sufficient to cause production of HA. The isolated nucleic acid sequences encoding REF1 is any suitable sequence of the invention as described above, and preferably is provided within an expression vector such that addition of the vector to an *in vitro* transcription/translation system results in expression of the polypeptide. The system further comprises the substrates for REF1, as previously described. Alternatively, the system further comprises the means for generating the substrates for an REF1 of the present invention. Such means include but are not limited to those previously described.

In other embodiments of the present invention, the methods for producing HA further comprise collecting the HA produced. Such methods are known generally in the art, and described briefly above. In yet other embodiments of the present invention, the HA are further purified, as for example is described in the Examples.

2. Using REF1 Polypeptides

In some embodiments of the present invention, methods for producing HA comprise incubating an HA of the present invention in an *in vitro* reaction mixture under conditions sufficient to result in the synthesis of HAs; generally, such incubation is carried out in a mixture which comprises substrates of REF1.

An REF1 of the present invention, as described above, is obtained by purification of either naturally occurring REF1 or recombinant REF1 from an organism transformed with heterologous gene encoding an REF1, as described above. A source of naturally occurring REF1 is contemplated to include but not limited to plants, as for example are listed in Table 2. A source of recombinant REF1 is either plant, bacterial or other transgenic organisms, transformed with heterologous gene encoding REF1 of the present invention, as described above. The recombinant REF1 may include means for improving purification, as for example a 6x-His tag added to the C-terminus of the protein as described above. Alternatively, REF1 is chemically synthesized.

The incubation mixture further comprises the substrates for REF1, as described above. Alternatively, the mixture further comprises the means for generating the substrates for REF1, such as coniferaldehyde or sinapaldehyde.

In other embodiments of the present invention, the methods for producing HA further comprise collecting the HA produced; such methods are described above.

EXPERIMENTAL

The following examples are provided in order to demonstrate and further illustrate certain preferred embodiments and aspects of the present invention and are not to be construed as limiting the scope thereof.

In the experimental disclosures which follow, the following abbreviations apply: N (normal); M (molar); mM (millimolar); μM (micromolar); mol (moles); mmol (millimoles); μmol (micromoles); nmol (nanomoles); pmol (picomoles); g (grams); mg (milligrams); μg (micrograms); ng (nanograms); l or L (liters); ml (milliliters); μl (microliters); cm (centimeters); mm (millimeters); μm (micrometers); nm (nanometers); °C (degrees Centigrade); abcissic acid (ABA); cleaved amplified polymorphic sequence (CAPS); ethane methyl sulfonate (EMS); expressed sequence tag (EST); open reading frame (ORF); untranslated region (UTR); aldehyde dehydrogenase (ALDH); coniferaldehyde dehydrogenase (CALDH); sinapaldehyde dehydrogenase (SALDH); caffeic acid/5-hydroxyferulic acid O-methyltransferase (COMT); caffeoyl CoA 3-O-methyltransferase

(CCoAOMT); cinnamoyl alcohol dehydrogenase (CAD); cinnamoyl CoA reductase (CCR); p-coumarate 3-hydroxylase (C3H); ferulate 5-hydroxylase (F5H); reduced epidermal fluorescence (ref); high pressure liquid chromatography (HPLC); gas chromatography-mass spectrometry (GC-MS);

EXAMPLE 1

Methods

This example describes methods used to clone, sequence, identify, and characterize an *REF1* gene and gene product.

A. Plant Material and Growth Conditions

Arabidopsis thaliana L. Heynh. was grown at a light intensity of 100 μE $m^{-2}sec^{-1}$ at 23°C under a photoperiod of 16 hr light/8 hr dark in REDI-EARTH potting mix (Scotts-Sierra Horticulture Products; Marysville, OH). All Arabidopsis used in the experiments were ecotype Columbia unless mentioned. Seven *ref1* alleles (*ref1-1* to *ref1-7*) were used in this study; except for *ref1*-3, which was isolated from a T-DNA mutagenized population in Wassilewskija (ws) ecotype, the rest were isolated from a M_2 screen of ethyl methanesulfonate-mutagenized plants as described in Ruegger and Chapple, 2001. All of the alleles used in this study were backcrossed at least two times to the corresponding wild type.

B. Cloning the *REF1* gene

The *ref1* mutant (Columbia background) was used in a cross to the Landsberg *erecta* ecotype to establish a mapping population. F1 individuals were allowed to self-pollinate, and F2 plants were screened for the *ref1* phenotype. Initial screening for the *ref1* phenotype indicated that the *ref1* phenotype is linked to the *glabrous* mutation on chromosome 3. Hence further experiments were not conducted to identify in which chromosome the *REF1* was located. DNA was extracted from the homozygous *ref1* mutants from F2 and F3 Columbia plants for use in PCR-based genotyping experiments. Individuals carrying recombinant chromosomes in the region of the *REF1* locus were used to determine a mapping interval for the *REF1* gene.

To determine whether the coding region of the putative *REF1* candidate gene (aldehyde dehydrogenase, At3g24503) was disrupted by a T-DNA insertion, DNA isolated from the *ref1-3* (T-DNA allele) and the wild type ws ecotype and subjected to polymerase

chain reaction (PCR) using oligonucleotides CC581 (5'-atgagaacggcaaatg-3') and CC582 (5'-ttacatccaaggggaattgtg-3'). The position of the T-DNA insertion within the *REF1* gene was further delineated using oligonucleotides CC587 (5'-ccacttctcatattcaacgac-3') and CC588 (5'-gtcgttgaatatgagaagtgg-3'). DNA from *ref1-3* plant was PCR amplified with oligonucleotides CC581 and T-DNA left border primer and sequenced to determine the exact location of the T-DNA integration in the *ref1*-3 mutant. Genomic DNA corresponding to the *REF1* coding region from the rest of the *ref1* alleles (*ref1-1* to *ref1-7*) was PCR amplified as two over lapping fragments using oligonucleotide combinations CC614 (5'-aatccactgcctttgctgac -3') / CC628 (5'-cggcgcgactcataagaa-3') and CC620 (5'-aattggagtggttggta -3') / CC632 (5'-agccgccttattatcattgg-3'). These DNA fragments were sequenced in both directions to identify mutations in the different *ref1* alleles.

C. Complementation of the *ref1* mutant

A BAC clone MIOB24 (gene bank accession AB020746) containing the *REF1* gene was obtained from Arabidopsis Biological Resource Center (ABRC). An 11 kb *Kpn* I-*Xho*1 DNA fragment containing the *REF1* gene was isolated from the BAC clone and cloned into pBluescript II (Stratagene, La Jolla, CA) plasmid. A 7414 bp *Csp*45 I DNA fragment comprising 3297 bp of the *REF1* promoter region, 3298 bp of the coding region, and 819 bp of the 3'-UTR was isolated and cloned into a *Cla* I site of pBluescript II to generate pCC619. A *Kpn* I – *Pst* I fragment from pCC619 containing the 7414 bp fragment of the *REF1* gene was cloned into *Kpn* I – *Pst* I site of pCAMBIA2300 binary vector to generate pCC621. Plasmid pCC623 was constructed by ligating a 5507 bp *Bam*H I fragment from pCC619 containing 1390 bp of the *REF1* promoter region, 3298 bp of the coding region, and 819 bp of the 3'-UTR into a *Bam* HI site of pCAMBIA 2300. The binary vectors (pCAMBIA 2300, pCC621 and pCC623) were introduced into *Agrobacterium tumefaciens* C58 pGV3850 (Zambrisky et al., 1983) by electroporation. The *ref1* mutants (*ref1-1*, *ref1-2* and *ref1-4*) were transformed by the floral dip method (Clough and Bent, 1998). The transformed plants were selected as described by Hemm et al., 2003. Three-week-old transgenic plants were analyzed visually under UV light and by HPLC for complementation of the sinapoylmalate phenotype.

D. Analytical methods

Sinapic acid, ferulic acid, *p*-coumaric acid, sinapoylmalate and feruloylmalate were extracted from tissues and separated by HPLC as described by Hemm et al., 2003.

Sinapoylcholine in seeds was analyzed as above, except that a PURESIL™ C18 column (Waters, Milford, MA; 1200 nm pore size, 5 μM particle size) was used. Cell walls were prepared and the cell wall bound phenolic esters were released and analyzed as described by Franke et al., 2002b. GC-MS analysis was performed as described by Franke et al., 2002a.

E. Protein Extraction and Enzyme Activity Assay

Plant material (1g) was ground in liquid nitrogen and extracted in 5 ml buffer containing 50 mM 4-(2-hydroxyethyl)-1-piperazine-ethanesulfonic acid (Hepes)-KOH, pH 8.0, 1 mM ethylenediaminetetraaacetic acid (EDTA), 5 mM dithiothreitol (DTT) and 10% glycerol (v/v) for 30 min at 4°C. The extract was centrifuged at 10,000 g for 20 min. The supernatant was collected and the protein precipitated by adding ammonium sulfate to 70% saturation. The precipitated protein was centrifuged and the pellet re-suspended in 1 ml extraction buffer. The protein was desalted using PD-10 SEPHADEX® G-25M column and used for the aldehyde dehydrogenase assay.

The aldehyde dehydrogenase assay was conducted in 200 μl volume containing 50 mM Hepes-KOH, pH 8.0, 5 mM DTT, 1 mM NAD^+, 100 μM of coniferaldehyde or sinapaldehyde and 10% glycerol (v/v). The buffer was pre-warmed to 30°C for 10 min and 10 μl of the enzyme extract was added to initiate the enzyme assay. The assay was conducted for 30 min at 30°C and stopped by adding 50 μl of 3 M acetic acid. 20 μl of the assay mixture was analyzed by HPLC as described by Hemm et al., 2003 to determine the amount of ferulic and sinapic acid produced.

F. REF1* Expression in *E. coli

The open reading frame of *REF1* was PCR amplified from an EST clone (gene bank ID - T43357) using oligonucleotide gaagaga<u>catatg</u>gagaacggcaaatg-3' and 5'-gata<u>ctcgag</u>catccaaggggaattgtg -3'. The oligonucleotides incorporated an *Nde* I site at the N-terminal end and an *Xho* I site at the C-terminal end of the open reading frame. The PCR product was digested and cloned into the corresponding sites of a pET30a vector (Novagen, Madison, WI) to generate pCC628. The recombinant plasmid was then introduced into *E. coli* BL21DE3 pLysS and used for protein induction. An *Xho* I fragment from pCC628 was sub-cloned into pBluescript II and sequenced. As a control pET30a vector was also introduced into *E. coli* BL21DE3 pLysS.

Overnight cultures of *E. coli* BL21DE3 pLysS containing pCC628 and pET30a vectors in Luria Bertani (LB) medium was diluted 50-fold and grown at room temperature (22°C) until the OD_{600} reached 0.7. The cells were then induced for 4h with 0.05 mM isopropyl-β-D-thiogalactopyranoside (IPTG). The cells were centrifuged and the bacterial pellets were washed in distilled water and re-suspended in 3 ml extraction buffer (50 mM Hepes-KOH, pH 8.0, 1 mM EDTA, 5 mM DTT and 10% glycerol v/v). The cells were lysed using a French pressure cell and centrifuged at 14,000 g for 30 min to remove cell debris. The supernatants were used for ALDH assay as described above.

EXAMPLE 2

Characterization and Identification of *REF1* Gene and Product

This example describes the results of experiments to clone, sequence, identify, and characterize an *REF1* gene and gene product.

A. *ref1* mutants accumulates reduced levels of sinapate esters

Seven independent alleles of *ref1* were originally identified in a mutant screen (Ruegger and Chapple, 2001). All of the *ref1* mutants were morphologically similar to wild type, but exhibited reduced levels of epidermal fluorescence when observed under UV light. Preliminary investigation of the *ref1-1* and *ref1-2* indicated that the sinapate ester content in leaf and mature seeds were only 30 % of the wild-type levels (Ruegger and Chapple, 2002). To determine the impact of *ref1* mutation on the sinapate ester content in the different alleles, sinapoylmalate levels in 3 week old leaves were analyzed. In all of the *ref1* alleles analyzed, the sinapoylmalate content was reduced to less than 10 % of the wild-type levels. The sinapoylcholine content in mature seeds of *ref1* plants was reduced to less than about 30 % of the wild type seeds. The apparent discrepancy in the levels of sinapoylmalate content reported previously (Ruegger and Chapple, 2001) and observed in these experiments suggests that the sinapoylmalate content in *ref1* plants may be developmentally regulated.

To further understand the developmental regulation of sinapate ester content in *ref1* plants, sinapoylmalate and sinapoylcholine content was determined at various ages of seedling and developing embryos of *ref1-1*, *ref1-2* and wild type plants. Sinapoylmalate levels in both wild type and *ref1* plants increased with age, except at 21 days when both

wild type and *ref1* plants showed a decrease in sinapoylmalate content as compared to the preceding stage. However, the overall sinapoylmalate content in *ref1* plants was less than 30% of the wild-type at a given seedling age. To determine the sinapoylcholine content in developing embryos, siliques from individual inflorescences were harvested in pairs from top to bottom when the bottom silique turned yellow. The top silique pairs represented the most immature embryos while the bottom silique pairs represented the most mature embryos. Sinapoylcholine content increased as the embryos matured in both ref1 and wild type plants. However, the *ref1* mutants accumulated only less than 30 % of the wild-type levels of sinapoylcholine during embryo maturation. These results suggest that the REF1 plays a substantial role in the accumulation of sinapate esters in both seedlings and embryos throughout its development.

B. ***REF1*** **gene encodes an aldehyde dehydrogenase**

To understand the molecular and biochemical basis of sinapate ester reduction in *ref1* plants, the *REF1* gene was isolated by a combination of map based and candidate gene approaches. Linkage analysis using a mapping population of 550 F2 plants showed that the *REF1* locus was positioned between cleaved amplified polymorphic sequence (CAPS; Konieczny and Ausubel, 1993) markers CER435669 and g4711 on chromosome 3. These two markers define a 154 kb region containing 39 annotated genes (Fig. 2). A CAPS marker between these two markers, CER438851, was in the zero recombination well, indicating that it is tightly linked to the *REF1* locus. The fully sequenced Arabidopsis genome was then used to identify candidate genes within this mapping region. One of these genes was annotated as a putative aldehyde dehydrogenase (ALDH) and was located less than 8 kb from the CAPS marker CER438851 at which no recombinants had been identified. To test the hypothesis that defects in this putative ALDH gene were responsible for the *ref1* phenotype, the *ref1-3* allele which had been generated by T-DNA mutagenesis was used. Although previous efforts using plasmid rescue to isolate the *REF1* gene with this allele had been unsuccessful, it was believed that this allele might have been the result of a partial/abortive T-DNA insertion event that would still be detectable by PCR. Indeed, PCR amplification using gene specific and T-DNA border primers showed that the putative ALDH gene was disrupted by a T-DNA insertion (Fig. 3).

Lack of PCR amplification of the putative ALDH gene could be the result of a deletion event in the 3'-region of the ALDH gene and DNA sequence downstream of the ALDH gene which could contain other annotated genes. The *ref1* mutation could be the

result of such a gene. To rule out this possibility, the steady state RNA level of the ALDH gene by RNA gel blot analysis was determined. In four of the six EMS-induced alleles, the level of ALDH gene expression was reduced, and in 2 of the alleles it was less than 40% of the wild-type. Given that transcript levels of alleles containing frameshift or nonsense mutations are frequently decreased (Hilleren and Parker, 1999), these data further suggest that the putative ALDH gene is a strong candidate for *REF1*.

To provide definitive evidence that the *REF1* gene is a putative ALDH, the ALDH genomic DNA corresponding to all of the EMS-induced alleles was sequenced. Sequence analysis showed that three of the *ref1* alleles had point mutations in the gene resulting in premature translational stop codons (*ref1-1* and *ref1-4*, Q344STOP; *ref1-2*, W452STOP; Fig. 3). A mis-sense mutation is present in *ref1-6* (G416R) and in *ref1-7* (G152E), and there is a nucleotide change in the intron 6 / exon 7 splice site junction (agGT-aaGT) that may lead to improper splicing in the *ref1-5* allele (Fig. 3).

To demonstrate that the *ref1* mutation is the result of a lesion in the aldehyde dehydrogenase gene, the *ref1* mutant was complemented with wild-type *REF1* gene. Two plasmids, pCC621 and pCC623, were constructed with 3.3 kb and 1.4 kb fragments, respectively, of the *REF1* promoter DNA sequences, the *REF1* coding sequence and the 819 bp of the 3'-untranslated region (Fig. 6A). These plasmids were transformed into *ref1* mutants (*ref1-1*, *ref1-2* and ref1-4) using the floral dip method (Clough and Bent, 1998). Transgenic plants, selected on kanamycin plates, were able to completely restore wild-type levels of sinapoylmalate in *ref1-1*, *ref1-2* and *ref1-4* plants (Fig. 6B). This provided unequivocal evidence that a mutation in the *REF1* gene was responsible for the *ref1* phenotype.

C. *REF1* belongs to a family of aldehyde dehydrogenase in Arabidopsis

An expressed sequence tag (EST) (gene bank # AA395226; clone ID 118C14XP) that was annotated to have the 5'-end of the *REF1* gene was obtained and sequenced completely; the sequences are shown in Fig. 4. The 1625 bp cDNA sequence (SEQ ID NO:1, Fig. 4A) contained a 43 bp- 5' untranslated region (UTR), 1506 bp of the predicted open reading frame (ORF) (SEQ ID NO:2, Fig. 4B) and a 76 bp of the 3'-UTR. The deduced polypeptide consists of 501 amino acids (SEQ ID NO:3, Fig. 4C) with a predicted molecular mass of 54.3 kD. Its sequence contains the conserved residues that have been identified in human liver ALDH as active site amino acids (Ni et al., 1997; Sheikh et al., 1997). The lack of N-or C-terminal extensions compared to the mammalian liver ALDH

suggests that the REF1 is likely to be tetrameric. Based upon the standardized ALDH nomenclature system proposed recently (Vasiliou et al., 1999), REF1 is a member of the ALDH2 family of aldehyde dehydrogenases.

A BLASTP search of the Arabidopsis database identified at least 14 other proteins that were annotated as aldehyde dehydrogenases. These include betaine aldehyde dehydrogenase, methylmalonate semialdehyde dehydrogenase, and an ABA-inducible aldehyde dehydrogenase (Kirch et al., 2001). The predicted amino acid sequences of these aldehyde dehydrogenases varied from 482 to 608 amino acids. The closest homologs to REF1 were At1g23800 and At3g48000, which showed 57 % identity at the amino acid level. Interestingly, REF1 is more closely related (65 to 77% identity) to ALDH sequences described as having "unknown function" from *Medicago truncatula*, *Zea mays*, *Glycine max*, *Oryza sativa*, and *Nicotiana tabacum*. Among the homologous sequences from other plants (homologs, or orthologs), maize mitochondrial aldehyde dehydrogenase RF2A, which had 56% identity to REF1, is implicated as a nuclear restorer of cytoplasmic male sterility (Liu et al., 2001; Skibbe et al., 2002). The identification of related sequences in other species indicates that REF1 is not unique to Arabidopsis.

Exemplary related REF1 sequences, identified by amino acid identity with AtRAF1, are shown in Fig. 5.

D. *ref1* mutants exhibit reduced levels of sinapaldehyde and coniferaldehyde dehydrogenase activity

Based upon the observations that *ref1* mutants showed significant reduction in both sinapoylmalate and sinapoylcholine content in leaves and seeds, respectively, as described above in Section A of this Example, it was hypothesized that REF1 converts sinapaldehyde to sinapic acid, which then forms the precursor to sinapate ester biosynthesis. To test this hypothesis, the specific activity of sinapaldehyde dehydrogenase (SALDH) activity was determined using desalted enzyme extracts prepared from wild-type, *ref1-1*, *ref1-2* and *ref1-5* plants. *In vitro* enzyme reaction with wild-type leaf extract produced a novel compound that co-chromatographed with authentic sinapic acid standard when analyzed by HPLC. GC-MS analysis of the same reaction product identified a novel peak whose retention time and mass spectrum precisely matched that of authentic sinapic acid standard which permitted the unequivocal identification of sinapic acid formation. Wild type leaf extracts showed SALDH activity with a specific activity of approximately 200 pkat mg^{-1}; however, *ref1* plants exhibited less than 20% of the wild-type levels of SALDH activity (Fig. 7).

These results suggest that the Arabidopsis *REF1* gene encodes an aldehyde dehydrogenase that plays a major role in the synthesis of sinapic acid.

Plants are known to accumulate cell-wall bound ferulic acid esters (Hartley and Ford, 1989). Based upon the observations described above, it was proposed that cell wall bound ferulate esters are synthesized from coniferaldehyde by an aldehyde dehydrogenase activity similar to the Arabidopsis REF1. This proposal was tested by analyzing the same enzyme extracts from wild-type and *ref1* leaves for coniferaldehyde dehydrogenase (CALDH) activity. As described above, the synthesis of ferulic acid by CALDH activity was identified by GC-MS. Wild type plant extracts exhibited CALDH activity, although with a slightly lower specific activity (approximately 135 pkat mg^{-1}) than SALDH activity. In *ref1* plants, the CALDH activity was also reduced to approximately 20% of the wild type plants (Fig. 7). These observation indicate that REF1 also contributes to the synthesis of ferulic acid in Arabidopsis. Further evidence demonstrating the role of REF1 in the synthesis of cell wall bound ferulic acid is described below.

E. *REF1* encodes a functional aldehyde dehydrogenase in *E. coli*

The evidence described above indicates that *REF1* encodes an aldehyde dehydrogenase. Moreover, in the Arabidopsis *ref1* mutants, the sinapaldehyde and coniferaldehyde dehydrogenase activity were substantially reduced. Taken together, these observations indicate that the *REF1* open-reading frame could encode a functional SALDH or CALDH. To demonstrate that the *REF1* gene encodes a SALDH and/or CALDH, the *REF1* open-reading frame was sub cloned into a pET vector and expressed in *E. coli*. *In vitro* enzyme assays were conducted with crude protein extracts from *E. coli* transformed with pET30a-*REF1* or with pET30a vector without the *REF1* open-reading frame as control. HPLC analysis of *in vitro* assay with crude enzyme extract from *E. coli* containing pET30a-*REF1* plasmid detected a novel peak that co-chromatographed with the authentic sinapic or ferulic acid standards when sinapaldehyde and coniferaldehyde were used as substrates, respectively. However, *E. coli* transformed with pET30a vector alone did not produce either sinapic or ferulic acid. These results conclusively demonstrate that the *REF1* gene is an aldehyde dehydrogenase with both SALDH and CALDH activity.

F. REF1 is needed for the accumulation of cell wall bound ferulic acid in wild-type plants and feruloylmalate in *fah1* mutants

The results described above demonstrate that CALDH activity in *ref1* plants was markedly reduced, and that the *REF1* open-reading frame encodes CALDH activity. If ferulic acid, the major cell wall linked hydroxycinnamic acid, is synthesized by REF1, then in *ref1* plants the amount of cell wall linked ferulate esters should be lower than the wild type plants. To evaluate this possibility, wild type and *ref1* cell wall preparations were subjected to alkaline hydrolysis to release cell wall bound esterified phenolics. HPLC analysis of cell wall hydrolysates revealed that in the *ref1* plants cell wall linked ferulate esters was reduced to approximately 50% that of the wild-type. Interestingly, *p*-coumaric acid, which is synthesized from cinnamic acid by the action of cinnamate 4-hydroxylase, was not reduced in the cell wall preparation from *ref1* plants. These results provide direct evidence that that the *REF1* gene plays a major role in the formation of cell wall linked ferulate esters.

The *fah1* mutant lacks sinapoylmalate, the major phenolic ester in leaf (Chapple et al., 2002); instead, this mutant accumulates trace amounts of feruloylmalate (Hemm et al., 2002). To test the hypothesis that REF1 is needed for the production of ferulic acid which serves as the precursor for feruloylmalate synthesis in *fah1* plants, the *ref1-fah1* double mutant was developed. If REF1 is needed for feruloylmalate synthesis, then in the *ref1-fah1* double mutant, feruloylmalate should be lacking or reduced. Soluble secondary metabolites which accumulated in different plants (wild-type, *ref1-7*, *fah1-2*, and *fah1* x *ref1* double mutants) were analyzed by HPCL; soluble leaf phenolics were first extracted in 50% methanol and then analyzed by HPLC. When the methanol soluble leaf extract was analyzed by HPLC, feruloylmalate was not detected in the *ref1-fah1* double mutant. This observation, along with the cell-wall bound ferulic acid content in *ref1* mutant, demonstrate that the REF1 activity is important for the formation of ferulic acid in Arabidopsis.

G. CALDH and SALDH activity is present in different plants

The results described above demonstrate for the first time that Arabidopsis REF1 enzyme extracts contain both coniferaldehyde dehydrogenase (CALDH) and sinapaldehyde dehydrogenase (SALDH) activity. Although sinapate esters are a hall mark of the family Brassicaceae to which Arabidopsis belongs, there are a wide range of plants that accumulate cell wall linked ferulate esters (Hatfield, 1993; Lam et al., 1996; Ralph et al., 1995). Thus, enzyme extracts from different plants were analyzed for both CALDH and SALDH activity,

to determine the primary route to ferulic and sinapic acid synthesis in most plants is from coniferaldehyde or sinapaldehyde, respectively, by aldehyde dehydrogenase activity. The plants included representative samples of dicot (Arabidopsis, tobacco and radish), monocot (maize), gymnosperm (pine) and pteridophyte (fern). *In vitro* enzyme assays using desalted crude enzyme extracts detected both CALDH and SALDH activity in all the plants analyzed (Table 2). The specific activities of SALDH and CALDH varied among the different plants assayed. Wild type Arabidopsis enzyme extract had higher SALDH and CALDH specific activity than did the *ref1* plant extract. Conversely, the ratio of SALDH to CALDH activity was above 1 for wild-type plant extract and less than 1 for *ref1* plant extract, which imply that in *ref1* plants CALDH activity was predominant.

Table 2. SALDH and CALDH activity in different plants.

Plant	**SALDH activity** pkat mg-1	**CALDH activity** pkat mg-1	**S/C**
Arabidopsis wild-type leaf	48.1 ± 0.9	35.3 ± 0.6	1.4
Arabidopsis wild-type leaf	9.3 ± 1.2	15.8 ± 1.5	0.6
Arabidopsis *ref1* leaf	8.5 ± 0.7	17.4 ± 0.7	0.5
Tobacco leaf	26.9 ± 1.7	19.5 ± 1.7	1.4
Maize leaf	34.2 ± 4.2	69.7± 9.5	0.5
Radish leaf	24.0 ± 1.2	65.9± 7.5	0.4
Pine xylem	8.7 ± 1.5	34.9 ± 3.8	0.2
Ceratopteris frond	79.4 ± 6.0	62.4 ± 3.5	1.3

All publications and patents mentioned in the above specification are herein incorporated by reference. Various modifications and variations of the described method and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the described modes for carrying out the invention which are obvious to those skilled in chemistry, and molecular biology or related fields are intended to be within the scope of the following claims.